International Journal of Engineering Business and Social Science
Vol. 3 No. 8, December 2025
e-ISSN: ISSN 2980-4108, p-ISSN: 2980-4272
https://ijebss.ph/index.php/ijebss

A Study on the Role of Item Response Theory in the Development of Computerized Adaptive Testing

Muhammad Gibran Alif Prasetya*, Arif Widiyatmoko, Dyah Rini Indriyanti, Novi Ratna Dewi
Universitas Negeri Semarang, Indonesia
Email: gibranalif45@students.unnes.ac.id* , arif.widiyatmoko@mail.unnes.ac.id, dyahrini@mail.unnes.ac.id, noviratnadewi@mail.unnes.ac.id

Keywords
computerized adaptive testing (cat), item response theory (IRT), systematic literature review, assessment efficiency, machine learning.

Abstract
The paradigm shift in educational assessment from conventional testing to Computerized Adaptive Testing (CAT) based on Item Response Theory (IRT) offers significant measurement efficiency but presents complexities related to validity and the integration of new technologies. A Study on the Role of Item Response Theory in the Development of Computerized Adaptive Testing aims to analyze trends, methodologies, and the role of IRT in the development and validation of CAT through a Systematic Literature Review (SLR) approach. Following the PRISMA protocol, data were collected from the Scopus database for publications from 2020 to 2025 and analyzed using an NVivo-assisted thematic approach on 12 selected articles. The results show an evolution of research focus from mere efficiency in reducing question items to the integration of multimodal data, such as machine learning and physiological signals for ability estimation. Although CAT has been shown to drastically improve test efficiency without reducing reliability, challenges related to test security (item preknowledge) and item bias (Differential Item Functioning) remain major obstacles. It is concluded that the future of CAT development depends on balancing algorithmic efficiency, multidimensional data integration, and system resilience to validity threats to create fair and transparent assessments.

INTRODUCTION

Assessments in the field of education and psychology have undergone a significant shift from conventional testing to Computerized Adaptive Testing (CAT), which is advanced thanks to the Item Response Theory (IRT) algorithm. These changes are driven by efficiency, allowing for a 50% reduction in test duration without compromising accuracy, and effectively addressing problems in participants with extreme abilities (Ayanwale & Ndlovu, 2024; Huda, et al., 2024; Şİmşek & TAVŞANCIL, 2022). The core of CAT calibration is the logistics IRT model (1PL-3PL) which is systematically real-time can estimate hidden capabilities (latent trait) and choose the most informative questions for each individual (Apró & Tajti, 2025; Stuttgart et al., 2023). However, the

implementation of CAT also faces challenges; The integration of Machine Learning (ML) and Large Language Models (LLM) is needed to increase transparency and Explainability system, given that traditional design is often a "black box" for stakeholders (Cheng et al., 2024; Over et al., 2024). In addition, there is a technical dilemma in balancing the consistency of ranking between participants with the accuracy of individual estimates, an issue that is often overlooked in conventional CAT design (Liu et al., 2024).

Adopt an approach Systematic Literature Review (SLR) has become crucial compared to single field research due to the massive fragmentation in the current CAT literature. Field studies tend to be isolated in specific contexts, such as cognitive diagnosis (CD-CAT) for math skills (Aşİret & SÜNBÜL, 2024; A. Li et al., 2021), vocational competency assessment (Novrianti & Sari, 2025), or non-cognitive psychological measurements such as personality and anxiety disorders (Linen et al., 2023; ŞİmŞek & TAVŞANCIL, 2022). A systematic review is needed to synthesize these innovative trends, especially in mapping how new grain selection methods, such as Maximum Deviation Global Discrimination Index (MDGDI) performs compared to traditional methods in balancing attribute coverage and grain exposure control (Demir & Gelbal, 2025; J. Li et al., 2021). SLR also allows for comparative evaluation of the effectiveness of different algorithms in different contexts, e.g. the use of Deep Reinforcement Learning versus Bayesian statistical methods for test length optimization (Fink) et al., 2025; Zoucha et al., 2024), which is not possible in a single empirical study.

The latest research developments for the period 2020–2025 show an intense scientific debate on integration Generative AI and Automated Item Generation (AIG) in the CAT ecosystem. Recent research highlights potential Cognitive Design System (CDS) and LLM to drastically reduce the cost of developing question banks through automatic item difficulty prediction, although this carries the risk of parameter uncertainty that needs to be mitigated (Luo & Yang, 2024; Russell-Lasalandra et al., 2024). On the other hand, the issue of fairness and Validity in AI-based assessment is in the spotlight, especially with regard to the detection of cultural and linguistic bias in large-scale adaptive tests such as PISA (Tang et al., 2024; Woo & Choi, 2025). Empirical facts also show that although CAT improves efficiency, there are concerns about item exposure rate on a limited question bank, which forces the development of new, safer selection algorithms (Ayanwale & Ndlovu, 2024; Huda, et al., 2024). In addition, recent studies have begun to explore the use of CAT for Benchmarking Models Machine Learning itself, extending the application of IRT beyond the traditional psychometric domain (Song & Flach, 2021).

A review of previous research reveals diverse findings but also significant methodological limitations. Studies such as Dwahdh & Alshraifin (2025) and Read et al., (2024) Finding inconsistencies in the test termination rules (Stopping Rules), where the use of Standard Error Single (SE) as a termination criterion often fails to guarantee equivalent precision on very short tests. The problem of "cold start" or lack of initial information about participants is also still a major obstacle affecting the accuracy of the initial estimates, despite mitigation efforts using collateral information and Machine Learning has begun to be researched (Kim & Yoo, 2024). Furthermore, the implementation of CAT in developing countries such as Indonesia faces unique infrastructure and digital literacy challenges, which affect the ecological validity of the developed systems (Huda, et al., 2024; Imawan et al., 2025). These studies are often limited to homogeneous or simulated populations Post-hoc, so that it is less able to capture the dynamics of participants' responses in situations High-Stakes the truth.

Research Systematic Literature Review, It is designed to fill these gaps by systematically synthesizing the literature, identifying patterns of findings, and exploring in depth the role of IRT in CAT development and validation. This review will analyze trends, methodologies, and key

findings from recent articles to identify best practices, unresolved challenges, and potential theoretical developments. The results of this SLR have high academic relevance for updating standards of validity in adaptive testing and practical relevance for policy developers in designing fair, efficient, and transparent grading systems in the era of digital transformation (Demir & Gelbal, 2025; Liu et al., 2024).

METHOD

This study used a systematic literature review (SLR) approach to map and evaluate the role of *Item Response Theory* (*IRT*) in the development of *Computerized Adaptive Testing* (*CAT*). Systematic studies were selected so that the process of identification, selection, and evaluation of the literature was carried out in a structured manner and could be replicated, while allowing the preparation of a comprehensive scientific synthesis. The research methodological framework was designed following the PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) standard, as commonly used in modern educational technology studies. In addition, this study utilized thematic analysis through NVivo to identify thematic relationships, the emergence of keywords, and mapping of author networks related to the development of IRT-based CAT. Thus, this study combined a systematic approach and an integrated thematic analysis to provide a comprehensive picture of the research landscape.

The main source of research data was obtained from the Scopus database, a leading scientific database that provides access to peer-reviewed literature across disciplines. The selection of Scopus was particularly relevant for the topic "*A Study on the Role of Item Response Theory in the Development of Computerized Adaptive Testing*" because of its broad scope in crucial fields such as educational technology, psychometrics, artificial intelligence, machine learning, and adaptive evaluation, all of which are closely related to *Item Response Theory* (*IRT*) and *Computerized Adaptive Testing* (*CAT*). Its high credibility, multidisciplinary scope, and compatibility with qualitative data analysis tools such as NVivo were the main reasons for choosing Scopus.

The data collection process was carried out directly through the Scopus website, which allowed for efficient searching, filtering, and downloading of article metadata. The metadata collected included a variety of important information such as author name, affiliation, year of publication, journal title, number of citations, and relevant keywords. To ensure relevance and timeliness, the search for documents was limited to the 2020 to 2025 publication year range, in order to capture the current phases of development and important integration of IRT in CAT development.

The literature selection process followed the PRISMA flow to ensure transparency in the identification, screening, and inclusion stages of documents. The initial search was performed on the Scopus database using the following Boolean query:

TITLE-ABS-KEY ("Computerized Adaptive Test" OR "CAT") and title-abs-key ("Item Response Theory" OR "IRT") and pubyear > 2019 AND PUBYEAR < 2026 AND NOT (DOCTYPE ("Conference Paper") OR DOCTYPE ("Review") OR DOCTYPE ("Chapter")) AND (EXCLUDE (LANGUAGE , "Chinese")

The search yielded a collection of documents relevant to IRT-based CAT research. In the next stage, language filters were applied to include only English-language publications so that terminological consistency could be maintained. In addition, document type screening narrowed the search results to only research articles, excluding short notes and publications that did not undergo a credible peer review process. This screening process ensured that the selected documents truly addressed the role of IRT in CAT development and validation. A PRISMA diagram depicting the stages of identification, screening, and inclusion was presented to demonstrate the transparency and reproducibility of this process.

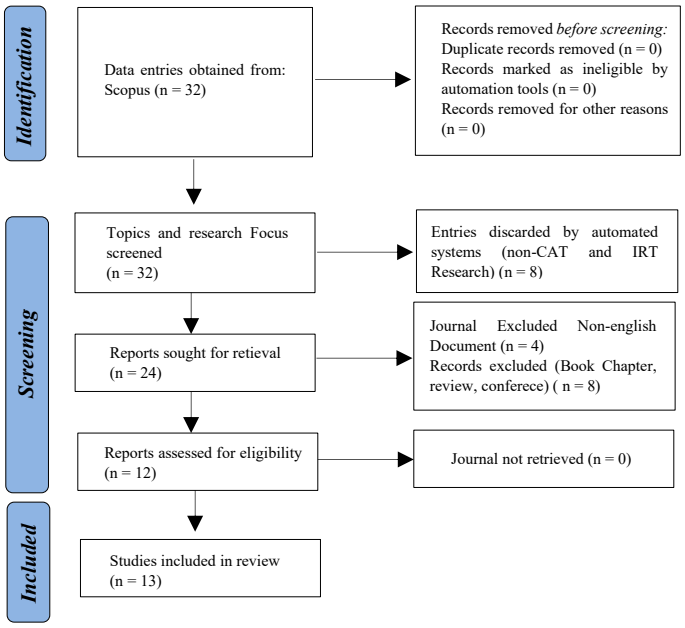

Figure 1. Article selection prism diagram

To ensure data completeness and support comprehensive thematic analysis, all search metadata from the Scopus database were exported in Research Information Systems (.ris) format. This format was chosen for its ability to accommodate rich bibliographic details, which were vital for analysis using NVivo. The next stage involved consolidating all exported RIS files, followed by a series of rigorous data cleansing processes. This process included identifying and eliminating duplicate entries, standardizing author name formats, harmonizing key terms and synonyms, and manually verifying article titles and abstracts. This manual verification was crucial to ensure the direct relevance of each article to the integration of Item Response Theory (IRT) in the context of Computerized Adaptive Testing (CAT), including discussions of item selection algorithms, ability estimation methods, and the application of varied IRT models. Once this series of processes was completed, a final processed dataset was formed, ready to serve as the foundation for qualitative and quantitative content analysis, as well as thematic mapping using NVivo.

To ensure the validity and reliability of the data in this study, a series of strict measures were implemented. Internal validity was ensured through careful cross-verification between the metadata extracted from Scopus and the full texts of the articles. This included confirming that the IRT models discussed, the algorithms used, and the reported CAT development methodologies aligned completely with the research focus. In addition, external validity was maintained by examining the consistency and coherence of thematic clusters resulting from qualitative analysis using NVivo, ensuring that the grouping of concepts was based on substantive relationships rather than metadata artifacts. Standardization of author names was also essential to prevent inaccurate fragmentation of the collaboration map.

Although the Scopus database is known for its broad scope and quality, some limitations were identified. These included potential variations in the depth of methodology descriptions between articles—especially related to the detailed implementation of IRT models or certain adaptive algorithms—as well as inconsistencies in keyword tagging by authors. However, these limitations were mitigated through a rigorous literature selection process, in which only studies meeting high academic standards and direct relevance to the research objectives were included. This approach

ensured that, despite the inherent limitations of the database, the research findings remained robust and reliable.

The analysis in this research was carried out using a mixed-methods approach that integrated qualitative and quantitative techniques. NVivo served as the main instrument for in-depth thematic analysis of the qualitative aspects, allowing the identification of key concepts, emerging themes, and relationships between ideas in relevant articles. The coding process was deployed to identify specific discussions of the IRT models used, the effectiveness of item selection algorithms, ability estimation methods, and various implementations of IRT in the context of CAT. In addition, for the quantitative aspects, frequency-based content analysis—such as the frequency of occurrence of certain keywords or relationships between codes—was conducted using NVivo features to present an initial picture of the dominance of important topics or trends. In more detail, each article was analyzed based on the IRT model applied (e.g., Rasch, 2PL, 3PL models), the type and effectiveness of adaptive item selection algorithms (e.g., Maximum Information, KL information), methods of estimating test participants' ability (e.g., Maximum Likelihood Estimation, Bayesian Estimation), variations of CAT developed for diagnostic, formative, or summative purposes, and its contribution to improving the efficiency, precision, and validity of adaptive testing. The combination of thematic and frequency analysis, both facilitated by NVivo, provided a comprehensive understanding of the role of IRT in CAT development, including methodological developments and relevant practical implications.

RESULTS AND DISCUSSION

Publication and Distribution Results of the Field of Study

Figure 2 presents the distribution of documents by subject area, which clearly shows the dominance of the fields of Psychology and Social Sciences, both of which account for the same largest portion at 20.6%, followed by Medicine with 19.0%; while the Arts and Humanities and Health Professions) accounted for the intermediate percentages of 7.9% and 6.3%, respectively. Other fields such as Computer Science, Mathematics, Neuroscience, and Other categories each contributed 4.8%, followed by Decision Sciences and Nursing) with the smallest portion of 3.2%, respectively, confirming the main focus of the document collection on psychometric foundations and social-medical applications, as shown in Figure 2.

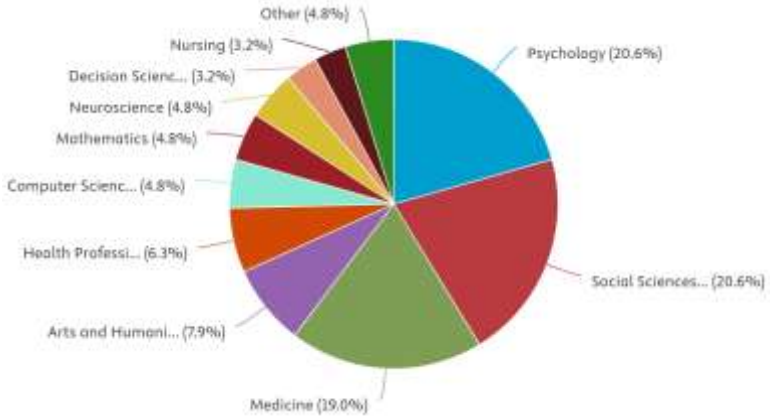

Figure 2. Documents by Subject Field

Author Productivity Analysis

The analysis in Figure 3 of the Documents by the Top 10 Authors shows an identical and even level of productivity among the leading researchers, where each author in the top 10 list,

including Cella, D., Erdem Kara, B., Fergadiotis, G., and so on to Terwee, C.B., each recorded contributing the exact same number of works, i.e. 2 documents; This uniform distribution pattern indicates that the development of literature on this topic is driven by a group of core researchers with an equal level of activity, in the absence of extreme single-dominated dominance of one particular individual in the data period taken, as shown in Figure 3.

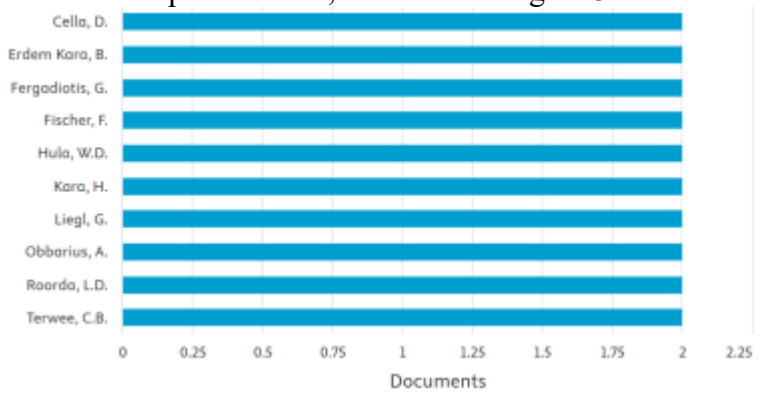

Figure 3. Documents by Top 10 Authors

Institutional and Affiliate Analysis

The document dataset also shows a specific pattern of academic concentration in terms of the affiliation of the institutions that publish or support it, as illustrated in Figure 4 presenting Documents by Top Institutional Affiliations, where Northwestern University Feinberg School of Medicine leads the contribution with the highest number of 5 documents, followed by Charité – Universitätsmedizin Berlin with 4 documents; Meanwhile, mid-tier institutions such as Northwestern University, Vrije Universiteit Amsterdam, and Amsterdam Public Health each contributed 3 documents, while other groups of institutions ranging from Humboldt-Universität zu Berlin to the University of Pittsburgh were recorded to contribute equally as many as 2 documents. This implies that the research on Item Response Theory and Computerized Adaptive Tests in this dataset has a very strong base in leading medical and public health institutions in the United States and Europe, indicating the intensive application of adaptive technologies in clinical or psychometric measurements of health, as shown in Figure 4.

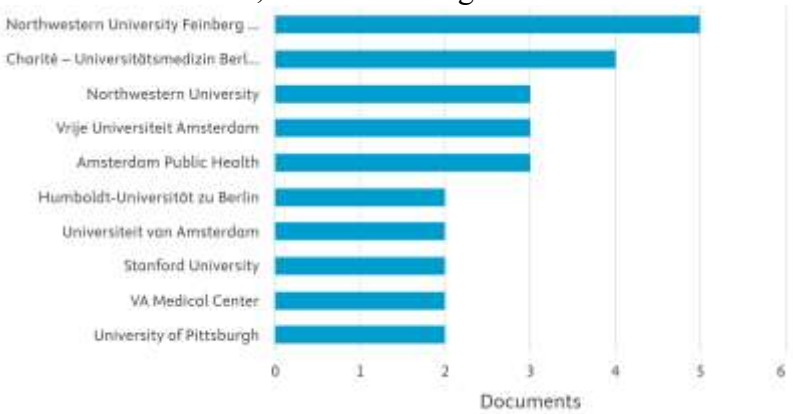

Figure 4. Documents based on Top 10 Publisher Parties

Data Visualization Results

The word cloud visualization in Figure 5, analyzed using the NVivo tool, clearly highlights the key keywords that are highly relevant to the title of this study. The keywords Item and Difficulty

appear very dominant, directly reflecting the core of Item Response Theory (IRT) which is centered on analyzing the characteristics of the item in question, especially the difficulty level of the item. In addition, words such as Measurement, Test, Scores, Model, and Methods also stand out, further strengthening the connection with IRT and the development of Computerized Adaptive Test (CAT). The presence of terms such as preknowledge, estimation, adaptive, responses, and learning indicates important aspects in the use and development of IRT/CAT, including the evaluation of initial knowledge, the estimation of item parameters and the ability of test participants, as well as the adaptive characteristics of the test that are able to adapt to individual performance. Based on NVivo's analysis, this word cloud confirms that this SLR research will closely interact with the core concepts in IRT and CAT, as well as the methodology used to develop and evaluate computerized adaptive tests.

Figure 5. Word Cloud

Figure 6 presents a tree map that, through NVivo analysis, visualizes the hierarchy and proportion of the occurrence of keywords in the corpus of this study. The size and position of the boxes in this tree map directly represent the frequency and relevance of the keywords. The large area dominated by the keywords Item, Difficulty, Test, and Based in the main cluster on the left, indicates the center of intensive scientific activity focused on the development of item selection algorithms and Computerized Adaptive Testing (CAT) performance evaluation based on Item Response Theory (IRT). Words such as change, measure, score, model, methods, and data also stand out, suggesting that these aspects are core to IRT studies and CAT development. Meanwhile, more specific terms with smaller box sizes on the right side of the tree map such as org, aberrant, data, preknow, learning, signals, journal, adapt, individ, system, know, task, informaa, indicate more specific or peripheral aspects of research. This visualization is essential to identify core and peripheral topics in the research domain, as well as help to understand how the literature develops cumulatively in a particular area so as to form different patterns of scientific concentration.

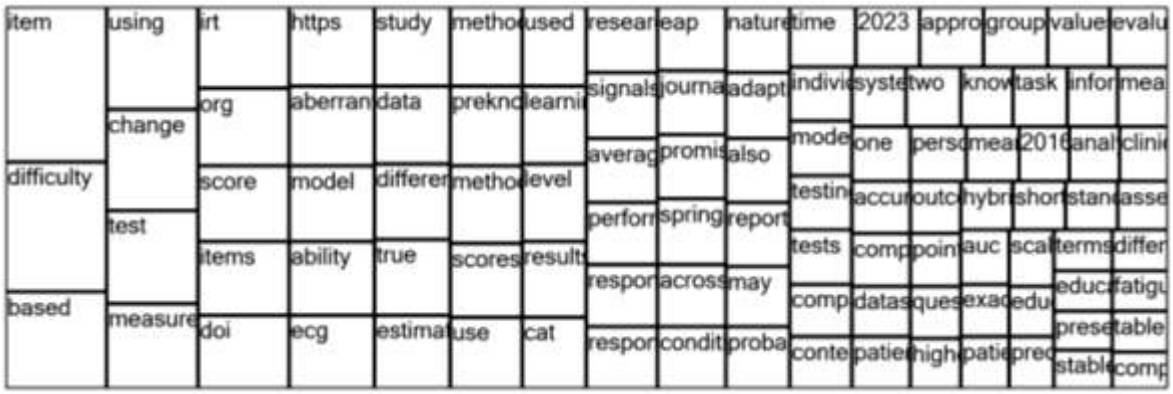

Figure 6. *Tree Map*

The cluster visualization in Figure 7 presents a thematic structure in the study "Study of the Role of Item Response Theory in the Development of Computerized Adaptive Test", where keywords are grouped based on co-occurrence and conceptual proximity. The dominant blue cluster in the top center represents the core of the topics of Item Response Theory (IRT) and Computerized Adaptive Test (CAT), encompassing fundamental technical keywords such as "item," "test," "model," "ability," "difficulty," "measurement," "score," "calibration," "parameters," "precision," and "estimation," which form the theoretical and practical foundations of this field. The brown/orange cluster at the bottom left highlights the specific applications and contexts of IRT and CAT, most likely related to aspects such as "education," "learning," "data analysis," "research," "study," "performance," "change," or specific research populations. The green cluster at the top right shows more methodological or evolving topics, such as "computer," "adaptive," "algorithm," "feedback," "system," "design," "efficiency," "validation," or "ethics" and "equity" issues in adaptive testing. Finally, small, scattered yellow/reddish clusters indicate specific issues, niche vocabulary, or concepts that are less prominent but relevant in a narrower context, such as specific statistics, analytics software, or unique implementation challenges. Thus, this visualization as a whole helps in understanding the conceptual landscape, key focus areas, as well as potential relationships between concepts in the literature.

Figure 7. *Cluster Analysis*

Qualitative Analysis of AI and IRT Integration in CAT Development to Improve the Effectiveness of Assessment in Learning

The following table presents a systematic synthesis of 12 selected articles illustrating the evolution of the role of Item Response Theory (IRT) in the Computerized Adaptive Test (CAT) ecosystem. Recent literature shows a paradigm shift in CAT development, moving from mere efficiency of reducing question items to the integration of more complex methodologies, such as the incorporation of physiological signals, machine learning, and response time estimation. This analysis maps how the IRT model is utilized not only as a statistical tool, but as an adaptive foundation for assessment innovation in the digital age.

Table 1. Qualitative Analysis Regarding the Role of IRT in CAT

No	References (Author's Name and Year)	The Role of IRT in CAT Development/Validation	Research Methodology	Key Findings & Contributions to SLR
1	Flag et al. (2024)	Assess the robustness of the estimated ability to test security threats (cheating).	Monte Carlo Simulation Study. Compare 2PL vs 3PL models and ML vs EAP estimates under preknowledge item conditions.	Security Challenges: EAP estimation methods are more resilient to fraud than ML. The 3PL model is recommended when the cheating rate is high to maintain the validity of the score
2	Arevalillo-Herráez et al. (2023)	Integrate IRT parameters with physiological data for real-time prediction of task difficulty.	Hybrid experiments. Fusion of IRT model with ECG signal data using machine learning (LDA) in language tests.	Theoretical Development: A hybrid approach (IRT + Biometrics) improves the accuracy of difficulty estimation compared to IRT alone. Demonstrate the evolution of IRT towards a multimodal adaptive system.
3	Kachergis et al., (2022)	Recalibrate the long instrument into a short adaptive test without loss of reliability.	Secondary Data Analysis & Simulation. 2PL IRT model on a large dataset (N>7000) for adaptive short-form development .	Best Practice (Efficiency): CAT with 25-50 question items is able to replace a questionnaire of hundreds of items with a high correlation (r>.95). New efficiency standards for child development assessment.
4	And et al. (2025)	Build a web-based reading assessment system that is self-sufficient (without proctoring) and efficient.	Software Development (jsCAT) & Empirical Validation. Validation in students in grades 1-8 compared the CAT randomly.	Accessibility & Scalability: ROAR-CAT increases efficiency by 40% and has high concurrent validity with oral tests. This model is relevant for large-scale assessment policies in schools.
5	Ince & Özbay (2025)	Expand the assessment model by including the response time variable	Algorithm Development & Empirical Studies. Integration of cubic/linear regression and	Assessment Innovation: Combining response times increases the sensitivity of

No	References (Author's Name and Year)	The Role of IRT in CAT Development/Validation	Research Methodology	Key Findings & Contributions to SLR
		along with the accuracy of the item.	AI (Random Forest) in the IRT framework.	students' ability classifications by up to 33%. Suggest new standards of fairer and time-aware judgment.
6	Ho et al. (2023)	Using IRT-based Plausible Values (PV) to measure individual clinical changes precisely.	Longitudinal Studies. Application of PV to PROMIS scores of COPD patients to detect true change vs measurement error	Longitudinal Validity Standard: CAT is more sensitive to detecting changes in patient conditions than short-forms. PV methods are important for distinguishing statistical and clinical changes in health monitoring
7	Liegl et al. (2023)	Validate the use of non-text item formats (video animations) in the CAT framework.	IRT Calibration & Post-hoc Simulation. Using the Graded Response Model (GRM) on the physical activity questionnaire (N = 1408).	Item Format Innovation**:** A valid video format is used in CAT, drastically reducing the burden of respondents (an average of < 8 items) with full instrument-equivalent precision.
8	Sahin Kursad & Yalcin (2024)	Evaluate the impact of item bias (Differential Item Functioning/DIF) on the integrity of the CAT system.	Simulation Studies. Manipulate percentages and DIF types to see their effect on measurement precision	System Fairness: The existence of DIF significantly impairs the precision and functionality of test information. Identification of DIFs is crucial for designing fair and unbiased assessment policies.
9	Xu (2024)	As a non-parametric comparator/alternative when parametric IRT assumptions are difficult to meet.	Development of new algorithms. Iterative Item Selection of Neighborhood Clusters (Non-IRT method) on personality tests.	Alternative Methodology: Offers solutions when complex IRT models cannot be implemented. Relevant for SLR as a comparison of IRT limitations in a given psychological context
10	Stuttgart et al. (2024)	Develop an item selection method for multidimensional tests with a forced-choice format.	Simulation Studies. The Thurstonian IRT model compares the Kullback–Leibler (KL) vs Fisher information method.	Theoretical Development: The KL-based selection method is superior for multidimensional forced-choice items , overcoming the weaknesses of the Fisher method in the early stages of the test. Important

No	References (Author's Name and Year)	The Role of IRT in CAT Development/Validation	Research Methodology	Key Findings & Contributions to SLR
				for the measurement of complex properties.
11	Komarc et al. (2024)	Refinement of the sexual knowledge scale for measurement efficiency in the young adult population.	Psychometric Analysis (CTT & IRT) & CAT Simulation. Student sample (N = 1291).	Efficiency & Validity: CAT reduces the number of grains by up to 54.3% without sacrificing reliability. Prove IRT's efficiency in filtering out bad items (low differentiation).
12	Apolone et al. (2024)	Implementation of CAT in multinational health policy instruments (cross-border standardization).	Validation Study Protocol. The Pan-European survey (N = 4500) used a static and dynamic toolkit (CAT).	Policy Relevance: An example of large-scale CAT implementation to address health inequalities in Europe. Demonstrate the role of IRT in creating transparent standards of assessment across cultures

A systematic analysis of the current literature shows that the application of Item Response Theory (IRT) in Computerized Adaptive Testing (CAT) has grown far beyond the efficiency of reducing question items, now expanding into multimodal data integration and ensuring test fairness. Consistently, the literature confirms that CAT is able to drastically improve measurement efficiency without sacrificing validity. This is evidenced by Kachergis et al., (2022) which succeeded in reducing children's vocabulary instruments from hundreds of items to only 25-50 items with high correlation ($r > 95$), and Komarc et al. (2024) which recorded a 54.3% reduction in the number of items on the student sexual knowledge scale while maintaining reliability. This efficiency is also validated in the context of basic education through the development of ROAR-CAT by And et al. (2025), which showed a 40% increase in efficiency compared to random-sequence tests, as well as in a clinical context through the use of video animation-based questionnaires that significantly reduced the burden on osteoarthritis patients Liegl et al. (2023).

In addition to efficiency, the main trend in the theoretical development of IRT today is the integration of external variables to improve the precision of capability estimation (Theta). Methodological innovations are seen in the combination of IRT parameters with physiological signals, such as electrocardiography (ECG), to predict the difficulty of the task real-time in intelligent learning systems (Arevalillo-Herráez et al., 2023). In line with that, İnce & Özbay (2025) Extend the scoring model by including the response time variable (Response time) using algorithms Machine Learning, which has been shown to increase the sensitivity of students' ability classifications by up to 33%. On the other hand, Stuttgart et al. (2024) enriching the realm of multidimensional measurement by developing information-based item selection methods Kullback–Leibler for formats forced-choice Using the model Thurstonian IRT, which is superior to conventional Fisher information methods. In comparison, Xu (2024) offers an alternative perspective by proposing a non-parametric algorithm (neighborhood clusters) which is effectively used when the strict assumptions of the IRT model are difficult to meet in psychological questionnaires.

However, the literature also highlights crucial challenges related to the validity, security, and fairness of scoring systems in the digital age. Flag et al. (2024) Through a simulation study, it was found that cheating in the form of prior knowledge of the question item (Preknowledge Item) can distort the estimation ability, where the estimation method Expected a Priori (EAP) and 3-Parameter Logistic (3PL) models have proven to be more resilient (Robust) compared to the Maximum Likelihood (ML) in mitigating these threats. Threats to validity also arise from the bias of question items, where Sahin Kursad and Sahin Kursad & Yalcin (2024) affirms that the existence of Differential Item Functioning (DIF) has a significant negative impact on the measurement accuracy and function of test information, demanding strict screening procedures to ensure fairness.

The practical relevance of these findings is particularly strong for large-scale health and education policy development. In the context of longitudinal health monitoring, Ho et al. (2023) demonstrate that the use of Plausible Values IRT-based on the PROMIS instrument is able to distinguish real clinical changes from mere measurement errors in chronic disease patients. The implementation of macro scale is clearly visible in the study protocol Apolone et al. (2024), which uses CAT in Toolkit EUonQoL to standardize the assessment of the quality of life of cancer patients in 25 European countries, proving that IRT-based CAT is a scalable solution to address health data inequities across countries. This synthesis confirms that the future of adaptive testing lies in the balance between algorithm efficiency, multidimensional data integration, and resilience to validity threats.

CONCLUSION

This Systematic Literature Review (SLR) on Item Response Theory (IRT) in Computerized Adaptive Testing (CAT) from 2020–2025 reveals a fundamental evolution from IRT as a statistical tool for item reduction efficiency (cutting test length by over 50% without compromising reliability) to hybrid integrations with AI, multimodal data (e.g., physiological signals, machine learning, response times), and robust estimation methods like Expected a Priori (EAP) over Maximum Likelihood (ML) for real-time ability precision in educational and clinical contexts. While enhancing transparency and scalability for cross-national assessments like global health monitoring, challenges persist in test security (item preknowledge) and fairness (Differential Item Functioning). The review maps CAT as a balanced system prioritizing algorithmic efficiency, data integration, and resilience to validity threats, offering an evidence-based understanding of adaptive assessments in the digital era. For future research, longitudinal studies could evaluate real-world CAT implementations in diverse cultural settings to address equity gaps and refine bias mitigation strategies.

REFERENCES

Apolone, G., Costantini, M., Caselli, L., Bos, N., Caraceni, A., Ciliberto, G., Couespel, N., & Ferrer, M. (2024). Validation of the European Oncology toolkit for the self - assessment of Quality of Life (EUonQoL - Kit) in cancer patients and survivors : study protocol of a pan European survey. *BMC Public Health*, *24*(3517). https://doi.org/10.1186/s12889-024-21008-4

Apró, A., & Tajti, T. (2025). An adaptive testing system for programming proficiency using Item Response Theory. *Annales Mathematicae et Informaticae*, *61*, 31–42. https://doi.org/10.33039/ami.2025.10.018

Arevalillo-Herráez, M., Katsigiannis, S., Alqahtani, F., & Arnau-gonzález, P. (2023). Fusing ECG signals and IRT models for task difficulty prediction in computerised educational systems.

Knowledge-Based Systems, *280*(July), 111052. https://doi.org/10.1016/j.knosys.2023.111052

Aşİret, S., & SÜNBÜL, S. Ö. (2024). Investigating The Performance of Item Selection Algorithms in Cognitive Diagnosis Computerized Adaptive Testing. *Journal of Measurement and Evaluation in Education and Psychology*, *15*(2), 148–165.

Ayanwale, M. A., & Ndlovu, M. (2024). The feasibility of computerized adaptive testing of the national benchmark test : A simulation study. *Journal of Pedagogical Reserach (JPR), 8*(2), 95–112.

Cheng, C., Zhao, G., Huang, Z., Zhuang, Y., Pan, Z., Liu, Q., Li, X., & Chen, E. (2024). Towards Explainable Computerized Adaptive Testing with Large Language Model. *Findings of the Association for Computational Linguistics: EMNLP*, 2655–2672.

Demir, H., & Gelbal, S. (2025). A Systematic Review on Computerized Adaptive Testing. *Erzincan University Journal of Education*, *27*(1).

Dwahdh, L., & Alshraifin, N. (2025). The impact of computerized adaptive test termination rules on accuracy across different ability estimation methods. *EURASIA Journal of Mathematics, Science and Technology Education*, *21*(1).

Fink, A., König, C., & Frey, A. (2025). Accounting for item calibration error in computerized adaptive testing. *Behavior Research Methods*. https://doi.org/10.3758/s13428-025-02649-8

Ho, E. H., Verkuilen, J., & Fischer, F. (2023). Measuring individual true change with PROMIS using IRT - based plausible values. *Quality of Life Research*, *32*(5), 1369–1379. https://doi.org/10.1007/s11136-022-03264-2

Huda, A., Firdaus, Irfan, D., Hendriyani, Y., Almasri, & 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.

Imawan, O. R., Retnawati, H., Haryanto, & Ismail, R. (2025). The challenges of implementing computerized adaptive testing in Indonesia. *Journal of Education and E-Learning Research*, *12*(2), 124–144. https://doi.org/10.20448/jeelr.v12i2.6677

Ince, A. H., & Özbay, S. (2025). AI-Enhanced Adaptive Testing : Integrating Response Time with IRT Models. *Ksii Transactions On Internet And Information Systems*, *19*(8), 2480–2498.

İnce, A. H., & Özbay, S. (2025). AI-Enhanced Adaptive Testing: Integrating Response Time with IRT Models. *KSII Transactions on Internet and Information Systems*, *19*(8), 2480–2498.

Kachergis, G., Marchman, V. A., Frank, M. C., Dale, P. S., & Mankewitz, J. (2022). Online Computerized Adaptive Tests of Children's Vocabulary Development in English and Mexican Spanish. *Journal of Speech, Language, and Hearing Research*, *65*(une), 2288–2308.

Kara, H., Dogan, N., & Kara, B. E. (2024). Robustness of Computer Adaptive Tests to the Presence of Item Preknowledge : A Simulation Study. *Journal of Measurement and Evaluation in Education and Psychology*, *15*(2), 138–147.

Kim, Y. J., & Yoo, J. E. (2024). Utilizing collateral information with machine learning in the cold-start problem in computerized adaptive testing: A Monte Carlo simulation study. *2024 IACAT Conference Program Book*.

Komarc, M., Shigeto, A., & Scheier, L. M. (2024). Item response theory and computer adaptive testing of the sexual knowledge scale of the sexual knowledge and attitude test in a college sample. *Psychology & Sexuality*, *15*(4), 661–678. https://doi.org/10.1080/19419899.2024.2332630

Lee, C. D., Han, K. C. T., Becker, K. A., Segall, D. O., Chang, H., Frey, A., Mead, A. D., & Wise, S. L. (2024). Evaluating the Effectiveness of the Standard Error of Score Estimation as a CAT Termination Criterion. *Journal of Computerized Adaptive Testing*, *11*(2). https://doi.org/10.7333/2410-1102013

Li, A., Sun, K., Wang, J., Wang, S., Zhao, X., Liu, R., & Lu, Y. (2021). Recombinant expression , purification and characterization of human soluble tumor necrosis factor receptor 2. *Protein Expression and Purification*, *182* (December 2020), 1–6. https://doi.org/10.1016/j.pep.2021.105857

Li, J., Ma, L., Zeng, P., & Kang, C. (2021). New Item Selection Method Accommodating Practical Constraints in Cognitive Diagnostic Computerized Adaptive Testing: Maximum Deviation and Maximum Limitation Global Discrimination Indexes. *Frontiers in Psychology*, *12*, 619771. https://doi.org/10.3389/fpsyg.2021.619771

Liegl, G., Roorda, L. D., Terwee, C. B., Steultjens, M., Roos, E. M., Guillemin, F., Grazia, M., Hanne, B., Carvalho, A. De, & Wilfred, B. (2023). Suitability of the animated activity questionnaire for use as computer adaptive test : establishing the AAQ - CAT . *Quality of Life Research*, *32*(8), 2403–2413. https://doi.org/10.1007/s11136-023-03402-4

Lin, Y., Brown, A., & Williams, P. (2023). Multidimensional Forced-Choice CAT With Dominance Items: An Empirical Comparison With Optimal Static Testing Under Different Desirability Matching. *Educational and Psychological Measurement*, *83*(2), 322–350. https://doi.org/10.1177/00131644221077637

Liu, Q., Yan, Z., Bi, H., Huang, Z., Huang, W., Li, J., Yu, J., Liu, Z., Hu, Z., Hong, Y., Pardos, Z. A., Ma, H., Zhu, M., Wang, S., & Chen, E. (2024). Survey of Computerized Adaptive Testing: A Machine Learning Perspective. *ArXiv*, *abs/2404.00712*. https://api.semanticscholar.org/CorpusID:268819592

Luo, H., & Yang, X. (2024). Efficiency of computerized adaptive testing with a cognitively designed item bank. *Frontiers in Psychology*, *15*, 1353419. https://doi.org/10.3389/fpsyg.2024.1353419

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*(56). https://doi.org/10.3758/s13428-024-02578-y

Novrianti, & Sari, L. C. (2025). Competence-Adaptive Assessment Using IRT for Junior Photographer Qualification. *Online Learning in Educational Research*, *5*(2), 305–317.

Russell-Lasalandra, L. L., Christensen, A. P., & Golino, H. (2024). Generative Psychometrics via AI-GENIE: Automatic Item Generation with Network-Integrated Evaluation. *Book of Abstracts - EAM2025*.

Sahin Kursad, M., & Yalcin, S. (2024). Effect of Differential Item Functioning on Computer

Adaptive Testing Under Different Conditions. *Applied Psychological Measurement*, *48*(7–8), 303–322. https://doi.org/10.1177/01466216241284295

Şİmşek, A. S., & Tavşancil, E. (2022). Applicability and Efficiency of a Polytomous IRT-Based Computerized Adaptive Test for Measuring Psychological Traits. *Journal of Measurement and Evaluation in Education and Psychology*, *13*(4), 328–344.

Song, H., & Flach, P. (2021). Efficient and Robust Model Benchmarks with Item Response Theory and Adaptive Testing. *International Journal of Interactive Multimedia and Artificial Intelligence*, *6*(5). https://doi.org/10.9781/ijimai.2021.02.009

Tang, X., Zheng, Y., Wu, T., Hau, K.-T., & Chang, Hua-Hua. (2024). Utilizing Response Time for Item Selection in On-the-Fly Multistage Adaptive Testing for PISA Assessment. *Journal of Educational Measurement*, *62*. https://doi.org/10.1111/jedm.12403

Wang, Q., Zheng, Y., Liu, K., Cai, Y., Peng, S., & Tu, D. (2024). Item selection methods in multidimensional computerized adaptive testing for forced-choice items using Thurstonian IRT model. *Behavioral Research Methods*, *56*(2), 600–614. https://doi.org/10.3758/s13428-022-02037-6

Woo, Y., & Choi, Y. (2025). Detection of cultural and linguistic differential item functioning in reading assessment. *Frontiers in Education*, *September*, 1–16. https://doi.org/10.3389/feduc.2025.1595658

Wulansari, A. D., Kirana, D. P., & Mufanti, R. (2023). Development of a Computerized-Adaptive Test to Measure English Vocabulary Size with IRT. *Indonesian Journal on Learning and Advanced Education*, *5*(3), 277–294. https://doi.org/10.23917/ijolae.v5i3.22953

Xu, Y. (2024). Iterative Item Selection of Neighborhood Clusters: A Nonparametric and Non-IRT Method for Generating Miniature Computer Adaptive Questionnaires. *Educational and Psychological Measurement*, *84*(2), 364–386. https://doi.org/10.1177/00131644231176053

Zoucha, J., Himelfarb, I., & Tang, N. (2024). Test Length Optimization with Deep Reinforcement Learning. *Book of Abstracts - Psychometric Society*.