Erzincan Üniversitesi Eğitim Fakültesi Dergisi
Erzincan University Journal of Education Faculty
2025 Cilt 27 Sayı 1 (137-150) https://doi.org/10.17556/erziefd.1577880
Derleme Makalesi / Review Article

A Systematic Review on Computerized Adaptive Testing*

Bireyselleştirilmiş Bilgisayarlı Testler Üzerine Sistematik Bir Derleme

Hümeyra Demir[1] **Selahattin Gelbal**[2]

[1] Res. Asst., Hacettepe University, Department of Educational Sciences, Ankara, Türkiye
[2] Prof. Dr., Hacettepe University, Department of Educational Sciences, Ankara, Türkiye

Makale Bilgileri

Geliş Tarihi (Received Date)

01.11.2024

Kabul Tarihi (Accepted Date)

20.02.2025

Sorumlu Yazar

Hümeyra Demir

Hacettepe Üniversitesi Eğitim Bilimleri Bölümü

hmeyrademir@gmail.com

Abstract: The aim of this research is to systematically review studies related to Computerized Adaptive Testing (CAT). Following systematic review guidelines, 110 articles were evaluated to seek answers to the established questions. These articles were analyzed based on their objectives, results, and recommendations, leading to a general conclusion. The compiled articles highlighted that innovative methods for CAT emerged as the most researched area. Within these innovative methods, the most studied topics were item selection algorithms and cognitive diagnosis computerized adaptive testing (CD-CAT). The results indicate that CAT enhances the accuracy and efficiency of tests through newly developed methods. It has been determined that CAT facilitates the provision of short and effective tests tailored to students' knowledge levels, ensures applicability across various disciplines, and offers the opportunity to reach large audiences through remote education platforms. The study concludes that to promote wider acceptance of CAT and increase its effectiveness, there is a need for the development of software tools and research focused on user attitudes. This study aims to identify potential future development areas for CAT, thereby enhancing the effectiveness of personalized assessment systems in education.

Keywords: Computerized adaptive testing, systematic review, innovative methods for CAT, research trends in CAT

Öz: Bu araştırmanın amacı, Bireyselleştirilmiş Bilgisayarlı Testler (BBT) ile ilgili yapılmış çalışmaları sistematik bir derleme yöntemiyle incelemektir. Sistematik derleme kurallarına uygun olarak, 110 makale detaylı bir şekilde değerlendirilmiştir. Bu makaleler, her birinin amacı, sonuçları ve önerileri doğrultusunda analiz edilerek genel bir çıkarıma ulaşılmıştır. Analiz sonucunda, BBT'ler için yenilikçi yöntemlerin en fazla araştırılan alan olduğu tespit edilmiştir. Özellikle, madde seçim algoritmaları ve bilişsel tanı tabanlı BBT'ler gibi konular, bu yenilikçi yöntemler arasında öne çıkmaktadır. Elde edilen sonuçlar, BBT'lerin yeni geliştirilen yöntemlerle testlerin doğruluğunu ve verimliliğini artırdığını göstermektedir. BBT'lerin, öğrencilerin bilgi seviyesine uygun, kısa ve etkili testler sunma konusundaki avantajları, farklı disiplinlerde uygulanabilirliği sağlama potansiyeli ve uzaktan eğitim platformları üzerinden geniş kitlelere ulaşma imkânı sunduğu belirlenmiştir. Ayrıca, BBT'lerin daha yaygın kabul edilmesi ve etkinliğinin artırılması için yazılım araçlarının geliştirilmesi ve kullanıcı tutumlarına yönelik daha fazla araştırma yapılması gerektiği sonucuna ulaşılmıştır. Bu çalışma, BBT'lerin gelecekteki potansiyel gelişim alanlarını belirleyerek, eğitimde kişiselleştirilmiş değerlendirme sistemlerinin etkinliğini artırmayı hedeflemektedir.

Anahtar Kelimeler: Bireyselleştirilmiş bilgisayarlı testler, sistematik derleme, BBT'ler için yenilikçi yöntemler, BBT'ler için araştırma eğilimleri

Demir, H. & Gelbal, S. (2025). A systematic review on computerized adaptive testing. *Erzincan University Journal of Education Faculty*, 27(1), 137-150. https://doi.org/10.17556/erziefd.1577880

Introduction

Computerized Adaptive Testing (CAT) has emerged as a prominent alternative to traditional paper-and-pencil assessments since the 1980s, driven by advancements in information technology and psychometrics (Chang & Ying, 2009). Unlike conventional tests that administer identical items to all examinees, CAT employs an algorithm that dynamically selects items based on individual responses, optimizing test length and enhancing measurement accuracy (Mead & Drasgow, 1993; Şenel, 2021). In this process, the examinee's ability estimate is continuously updated after each response, guiding the selection of subsequent items—correct responses lead to more challenging items, while incorrect responses result in easier items being administered (Meijer & Nering, 1999; Kingsbury & Zara, 1989; Van der Linden & Glas, 2002). For the adaptive process to function effectively, the design of CAT necessitates several fundamental components. Reckase (1989) identified four key components that form the foundation of CAT's operation: the item pool, the item selection method, the ability estimation procedure, and the termination criterion. Together, these components constitute the sequential algorithm that ensures the test is optimally tailored to each individual examinee.

The item pool plays a central role in providing adequate information for participants across different ability levels. Maintaining its quality requires continuous updates—replacing obsolete items with newly developed ones to reflect changes in educational standards and societal contexts. These updates can be conducted through traditional calibration, where new items are field-tested alongside existing ones, or through online calibration, which enables real-time adjustments for greater efficiency (Kang et al., 2020). Moreover, maximizing the potential of these diverse CAT applications necessitates effective management of the item pool. Ensuring long-term test validity and reliability involves controlling item exposure, monitoring item drift, and maintaining content balance (Leroux et al., 2019; Weiss & Şahin, 2024). Item exposure control techniques prevent the overuse of frequently administered items while promoting the selection of underutilized ones, thereby enhancing test security (Bock et al., 1988). Additionally, item drift—which refers to changes in item parameters over time due to cultural and educational shifts—must be closely monitored to ensure the

* Bu çalışma birinci yazarın ikinci yazar danışmanlığında hazırladığı doktora tezinin bir bölümünden yararlanılarak oluşturulmuştur.

comparability of scores (Chen et al., 2003). Finally, content balancing ensures fair representation of all content areas within the test, preventing bias and maintaining psychometric integrity (Chen et al., 2003).

While item pool management plays a critical role in enhancing the accuracy and reliability of CAT, the overall performance of the system also depends on other key components, such as item selection algorithms, ability estimation techniques, and test termination criteria, all of which must be addressed through a comprehensive approach. The item selection method ensures that the most suitable items are administered based on the respondent's ability, with the Maximum Fisher Information (MFI) method being one of the most commonly used approaches (Şenel, 2021). Ability estimation considers the correctness of responses and item parameters, employing methods such as Maximum Likelihood Estimation (MLE) and Bayesian-based techniques (Hambleton et al., 1991; Şenel, 2021). The test terminates based on predefined rules, such as reaching a fixed number of items, achieving the desired measurement precision, or reaching the allotted time limit (Segall, 2005).

Advancements in technology and measurement theory have significantly expanded the applications of CAT. Among these, Multistage Testing (MST), Cognitive Diagnostic CAT (CD-CAT), Computerized Adaptive Classification Tests (CACT), and Multidimensional CAT (MCAT) are particularly noteworthy. While MST offers personalization at the module level rather than the item level, CD-CAT integrates cognitive diagnosis with adaptive testing to provide detailed insights into students' strengths and weaknesses. CACT focuses on classifying participants into predefined ability groups, terminating the test once a classification decision is reached. MCAT, based on Multidimensional Item Response Theory (MIRT), measures multiple traits simultaneously and can apply either compensatory or non-compensatory models, depending on the desired testing approach (Chang, 2019; Jodoin et al., 2006; Lin & Hsu & Wang, 2019).

Despite these advancements, a critical review of the literature reveals gaps in the understanding and implementation of CAT, particularly in recent years. There is a notable lack of comprehensive studies that systematically examine the objectives, findings, and recommendations of CAT research conducted in the past five years. Therefore, this study aims to compile a systematic review of the current research on computerized adaptive testing. The analysis will focus on identifying key gaps, providing guiding recommendations for future research, and raising awareness of the significance of personalized approaches in measurement and evaluation. Ultimately, this review seeks to contribute to academic research while addressing practical issues in CAT applications, supporting the development of more reliable, valid, and efficient adaptive assessments.

Based on this information, the purpose of this study is to conduct a systematic review of research on computerized adaptive testing to identify the current state and future research needs. Specifically, this study aims to address the following questions:

1. How can these studies be classified based on their objectives?
2. What findings related to CAT have been reported in these studies?
3. What recommendations for future research on CAT have been provided in these studies?

Using predefined criteria, a systematic review has been conducted to answer these questions, highlighting the similarities and differences among studies related to CAT.

Method

This study was conducted following the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Page et al., 2021). The systematic review aimed to analyze and structure existing literature on CAT.

Eligibility Criteria

All studies investigating CAT were eligible for inclusion in this systematic review. The inclusion criteria were: (i) studies published between 2019 and 2024, (ii) written in English, (iii) peer-reviewed journal articles, (iv) focused on the social sciences, and (v) directly related to CAT applications.

Studies were excluded if they met any of the following criteria: (i) mentioned CAT but were not directly focused on its applications, (ii) not published in English, or (iii) were conference papers, book chapters, or other non-peer-reviewed formats. Figure 1 presents the inclusion and exclusion criteria applied in this systematic review. The figure visually summarizes the criteria used to select relevant studies and to exclude those that did not meet the predefined conditions.

Information Sources and Search

This study conducted a systematic search in the Scopus database to identify relevant studies. Scopus was selected as the sole database due to its unrestricted access, extensive content coverage, ease of use, and practicality. Its impact indicators are also considered more reliable and less prone to manipulation compared to those provided by WOS (Pranckutė, 2021). Another reason for choosing Scopus is its comprehensive coverage of peer-reviewed journal articles in the social sciences, which aligns well with the objectives of this research. The search was limited to studies published between 2019 and 2024 to maintain relevance and ensure the inclusion of the most recent research in the field.

Inclusion Criteria	Exclusion Criteria
Published between 2019 and 2024	Studies published outside the 2019-2024 range
Written in English	Not published in English
Peer-reviewed journal articles	Conference papers, book chapters, or other non-peer-reviewed formats
Focused on the social sciences	Studies from other subject areas (e.g., engineering, medicine, natural sciences)
Directly related to CAT	Studies not related to CAT or not directly focused on CAT

Figure 1. Inclusion and exclusion criteria

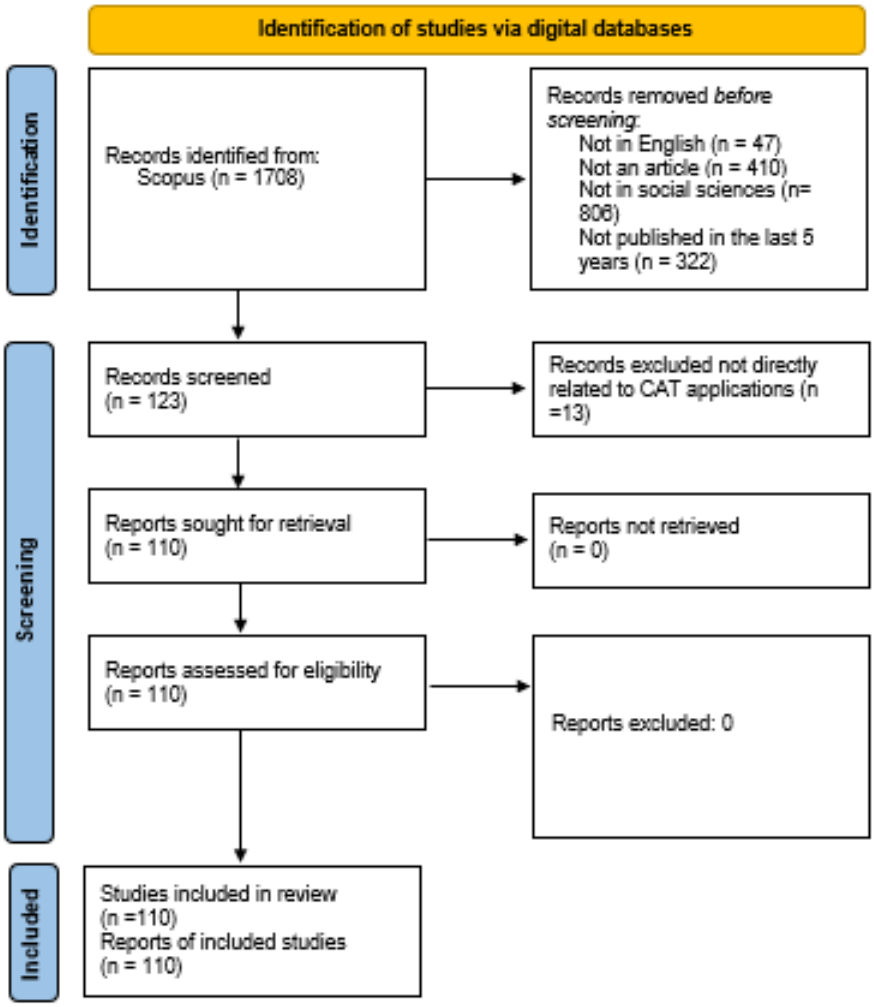

Figure 2. PRISMA flowchart for this study's methodology

The keyword 'computerized adaptive testing' was used for the search, targeting titles, abstracts, and keywords. Although various terms exist in the literature, 'computerized adaptive testing' was selected as it is the most general and widely recognized term. Moreover, it was observed that a significant proportion of the studies meeting the inclusion criteria incorporated this term, thereby supporting its exclusive use in our search strategy.

As shown in Figure 1, the initial search yielded 1,708 records. Inclusion and exclusion criteria were then applied to refine the results. Articles not written in English (n = 47), those not categorized as journal articles (n = 410), those outside the social sciences (n = 806), and those not published within the last five years (n = 322) were excluded at this stage. After these exclusions, 123 records remained for further screening.

During the screening process, 13 articles were excluded as they were not directly related to CAT applications. Consequently, 110 studies met all the inclusion criteria and were included in this systematic review. The PRISMA flowchart was prepared following the guidelines of systematic reviews suggested by Page et al. (2021) to ensure methodological transparency and clarity.

Findings and Discussion

The articles included in the review were analyzed in detail in line with the research questions and presented systematically. These articles were examined in terms of their objectives, findings, and recommendations, with each aspect discussed under separate headings in this section.

Figure 3 presents the distribution of articles included in the systematic review according to their publication years. As shown in the figure, the number of articles fluctuates between 2019 and 2024. The highest number of articles was published in 2022, with 21 articles, indicating increased interest in CAT during that year. On the other hand, the lowest number of articles was observed in 2024, reflecting a possible delay in publication processes for recent research.

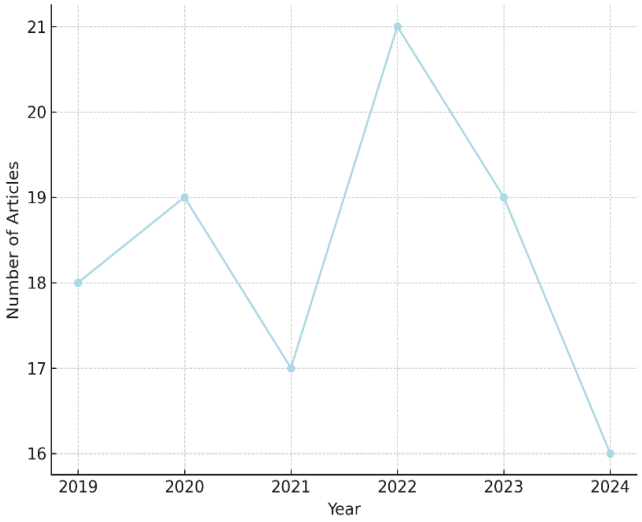

Figure 3. Distribution of articles by year

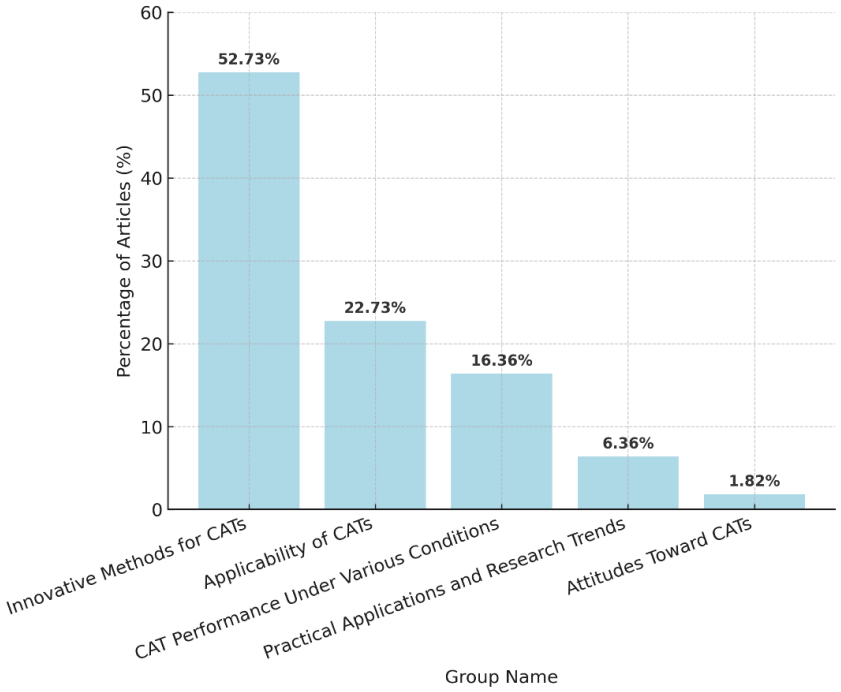

Figure 4. Grouping of CAT studies based on their objectives

Findings and Discussion Based on the Objectives of the Reviewed Studies

Studies on CAT were categorized based on their objectives, and these studies were grouped under five main categories. The categorization process was conducted collaboratively by both authors to ensure accuracy and reliability. The second author served as an additional expert to validate the categorization, and the process was carried out through consensus by independently reviewing and discussing the classification of studies. The distribution of these five categories and the number and percentages of articles in each group are presented in Figure 4.

Studies focusing on the development of new and innovative methodologies to improve CAT processes are categorized under the heading "Innovative Methods for CATs." This group includes studies proposing new item selection algorithms, ability estimation methods, test termination strategies, item pool optimization techniques, Differential Item Functioning (DIF) detection methods, online calibration techniques, and other technical advancements. These studies generally offer innovative solutions aimed at improving the accuracy and efficiency of tests. Among the 110 studies reviewed, 58 (52.73%) fall under this category.

Within the studies proposing innovative methods, the most frequently studied topics are item selection and CD-CATs. Specifically, 39.66% of the studies in this group propose new item selection methods for CATs (Bengs et al., 2021; Braeken & Paap, 2020; Chen & Chao, 2024; Chen C.W. et al., 2020; Davison et al., 2023; Gu et al., 2019; He & Qi, 2023; Hsu & Wang, 2019; Hsu & Wang, 2022; Kang et al., 2024; Kárász et al., 2023; Lin & Chang, 2019; Pan et al., 2023; Qiu et al., 2022; Tang et al., 2024; Wang, 2021; Xi et al., 2022; Xiao & Bulut, 2022; Yang et al., 2020; Yiğit et al., 2019; Yuan et al., 2023; Yuhana et al., 2024). These studies aim to enhance the efficiency of testing processes by improving item selection algorithms to make CAT processes more effective. Figure 5 presents an overview of the main research focus areas in studies proposing innovative methods for CATs.

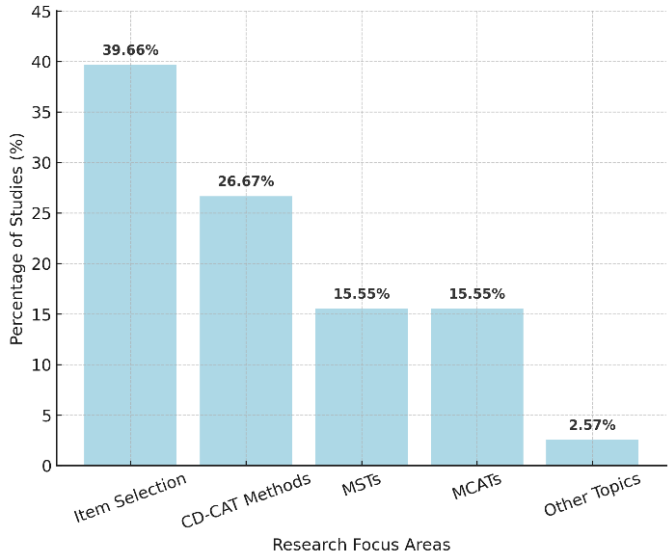

Figure 5. Distribution of innovative methods for CATs

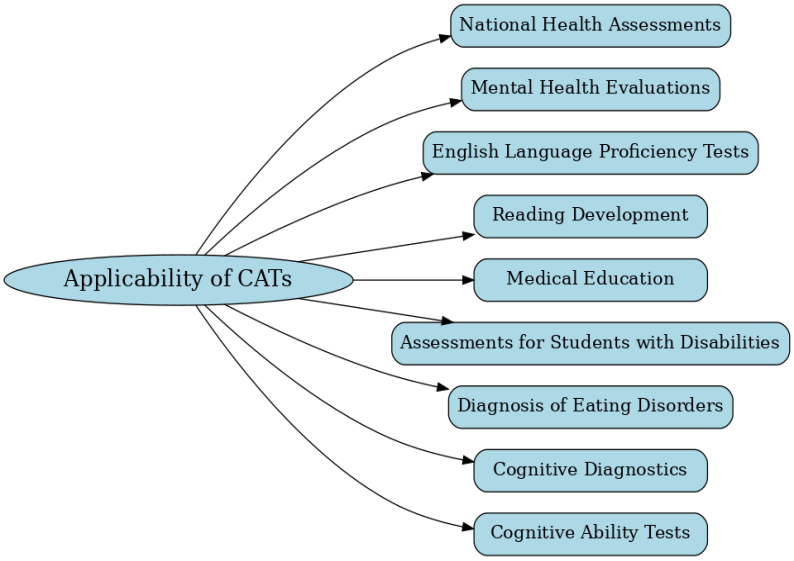

Figure 6. Classification of studies in the applicability of CATs category based on their objectives

Additionally, 26.67% of the studies propose innovative methods related to CD-CATs (Gao et al., 2020; Hsu & Wang, 2022; Kaplan & De La Torre, 2020; Lin & Chang, 2019; Luo et al., 2022; Sun et al., 2021; Tan et al., 2023; Wang, 2021; Wang et al., 2020; Yang et al., 2020; Yiğit et al., 2019). These studies address topics such as improving item selection algorithms in CD-CATs, reducing misclassification costs, increasing the flexibility of CD-CATs, working with various item types, and enhancing measurement precision and reliability.

In the "Innovative Methods for CATs" category, 15.55% of the studies focus on MSTs (Frey et al., 2023; Han, 2020; Kim & Yoo, 2023; Luo & Wang, 2019; Raborn & Sarı, 2021; Tang et al., 2024; Xiao & Bulut, 2022; Yang & Reckase, 2020), while another 15.55% concentrate on MCATs (Braeken & Paap, 2020; Chen et al., 2020; Luo et al., 2022; Qiu et al., 2022; Wang et al., 2022; Yuan et al., 2023). Other innovative approaches within this category include the creation of parallel item pools, optimal item pool design, and the development of item pool quality indices (Gönülateş, 2019; Lim & Han, 2024; Yang & Reckase, 2020); the incorporation of response time in item selection (He & Qi, 2023; Kang et al., 2020; Kern & Choe, 2021; Tang et al., 2024); and controlling item exposure rates (Chao & Chen, 2023; Chen et al., 2020; Chen et al., 2020; Pan et al., 2023; Qiu et al., 2022; Yasuda et al., 2022).

Moreover, studies have explored how the inclusion of response times (He & Qi, 2023; Kang et al., 2020; Kern & Choe, 2021; Tang et al., 2024) and response revisions (Lin et al., 2021; Wang et al., 2021) in CATs can enhance testing efficiency. Controlling item exposure rates has been considered a critical component for ensuring a fair testing experience (Chao & Chen, 2023; Chen et al., 2020; Chen et al., 2020; Pan et al., 2023; Qiu et al., 2022; Yasuda et al., 2022).

Research on online calibration methods (Kang et al., 2020; Tan et al., 2023; Yuan et al., 2023) aims to make multidimensional tests more effective and accurate. These studies seek to dynamically and in real-time improve online calibration processes by modeling the relationships between item responses and response times. Lastly, this category also includes studies introducing new methods for detecting DIF in CATs (Gu et al., 2019; Lim & Choe, 2023; Wang & Zhu, 2024). These studies propose methods for identifying items with DIF and aim to ensure fairness in CATs.

The second category, titled "Applicability of CATs," examines studies that explore how CATs can be utilized across various domains and applications. These studies focus on the applicability and validity of CAT implementations in specific areas or groups. The success of adaptive tests has been thoroughly analyzed, particularly in fields such as education, healthcare, and psychology. Of the 110 studies reviewed, 25 (22.73%) fall under this category. These studies investigate the applicability of CATs across a broad range of fields and populations, including national health assessments, mental health evaluations, English language proficiency tests, reading development in primary school students, medical education, assessments for students with disabilities, diagnosis of eating disorders, cognitive diagnostics, and cognitive ability tests. Figure 6 presents a classification of studies in the Applicability of CATs category.

Furthermore, the studies address the potential use of CATs in evaluating university students' mathematics knowledge levels, adapting measurement tools such as the Force Concept Inventory and the MacArthur-Bates Communicative Development Inventories into a CAT format, measuring high-level thinking skills in high school physics classes, and reducing mathematics test anxiety. These findings highlight the versatility of CATs across diverse areas and target groups (Adams et al., 2024; Ayanwale & Ndlovu, 2024; Chai, Lo, & Mayor, 2020; Ebenbeck & Gebhardt, 2022; Ebenbeck & Gebhardt, 2024; Ghio et al., 2022; Huang et al., 2022; Istiyono et al., 2020; Kaplan & De La Torre, 2020; Kaya et al., 2022; Komarc et al., 2024; Li et al., 2023; Liu et al., 2022; Norfarah et al., 2019; van Wijk et al., 2024; Yasuda et al., 2021).

The third category, titled Performance of CATs Under Various Conditions, encompasses studies that investigate the performance of CATs under different methodological conditions. These studies examine the impact of variables such as test length, item pool size, item selection algorithms, content balance, and ability estimation methods on the performance of CATs. Of the 110 studies reviewed, 18 (16.36%) fall under this group, with a focus on optimizing CAT performance using different parameters and algorithms. The studies aim to enhance test accuracy, optimize test durations, and ensure more efficient utilization of item pools. Figure 7 presents a classification of studies in the Performance of CATs Under Various Conditions category.

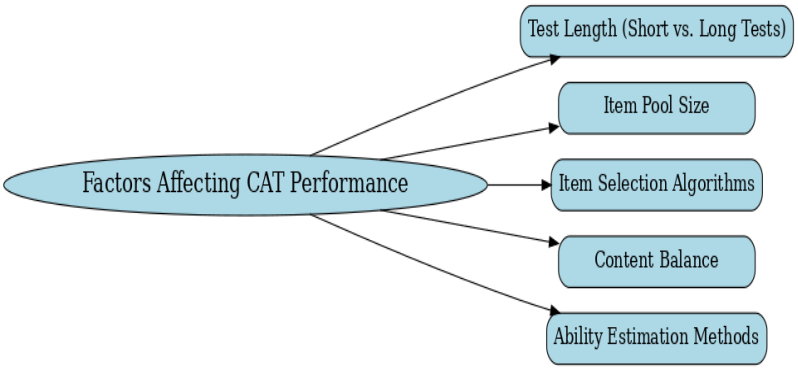

Figure 7. Classification of studies in the performance of CATs under various conditions category based on their objectives

For example, one study examined the effects of item selection algorithms on measurement accuracy and computational efficiency in Cognitive Diagnostic CATs (CD-CATs) under varying test lengths and item quality conditions (Aşiret & Sümbül, 2024). Another study explored the effects of factors such as item exposure rates and content balance on ability estimation accuracy, comparing different ability estimation methods under various conditions (Giray & Kelecioğlu, 2024). A different study analyzed how parameter variations in automatically generated items impact person parameter estimates and whether these variations have a balancing effect in longer tests (Tian & Choi, 2023). Other research investigated CATs with forced-choice items under adaptive and static conditions, focusing on social desirability balance and examining how adaptive item selection affects measurement accuracy in short and long tests (Brown & Williams, 2023). Additional studies compared the effectiveness of different adaptive testing approaches under varying test lengths and ability estimation methods, analyzing how test length and estimation methods influence performance.

The studies also focused on the effects of shadow testing methods on measurement accuracy, the impact of item exposure rate control on CAT performance, the influence of item selection methods on performance in item pools of different dimensions, and how item pool characteristics affect ability estimation and item utilization. These works collectively aim to optimize CAT performance by employing diverse algorithms, test structures, and methodological parameters, thereby improving test accuracy, optimizing durations, and enhancing the efficiency of item pools (Aşiret & Sünbül, 2024; Cooperman et al., 2022; Lin et al., 2023; Mao et al., 2022; Özdemir & Gelbal, 2021; Öztürk & Şahin, 2019; Sulak & Kelecioğlu, 2019; Tian & Choi, 2023; Tseng, 2021; Wyse, 2021; Yıldız et al., 2024; Yiğiter & Doğan, 2024).

The fourth category, Practical Applications and Research Trends, focuses on the applications of CATs in various fields, software tools developed for CATs, and the overall trends in research on CATs. This group highlights the practical use of CATs in areas such as licensing exams for healthcare professionals and educational assessments, alongside the features of software packages and technical tools employed in these contexts. Furthermore, this category includes literature analyses that examine prominent trends and methodological approaches in CAT-related research. Figure 8 presents a classification of studies in the Practical Applications and Research Trends category.

For example, Yurtçu and Güzeller (2021) conducted a bibliometric analysis of CAT-related research, while Lim and Wells (2020) introduced an R package that facilitates online item calibration and model fit evaluation. Yiğiter and Doğan (2023) provided guidance on the application methods of CATs using principles, designs, and simulations based on the MST framework in the R programming language. Meanwhile, Seo and Choi (2020) introduced a web-based CAT platform and examined its potential for use in healthcare, licensing, and certification exams.

The fifth category, Attitudes Toward CATs, focuses on studies examining the attitudes, perceptions, and acceptance levels of individuals and institutions toward the use of CATs. Only two (1.82%) of the 110 reviewed studies fall into this category, focusing on factors such as anxiety, motivation, and general perceptions toward CATs. These studies evaluate the attitudes of students and educators toward CATs, particularly the acceptance levels of medical students and academic staff and their views on the future adoption of adaptive tests. Additionally, the studies analyze whether CATs are perceived as a transformative tool in education and assessment processes (Kisielewska et al., 2024; Pramjeeth & Ramgovind, 2023). Figure 9 presents a classification of studies in the Attitudes Toward CATs category.

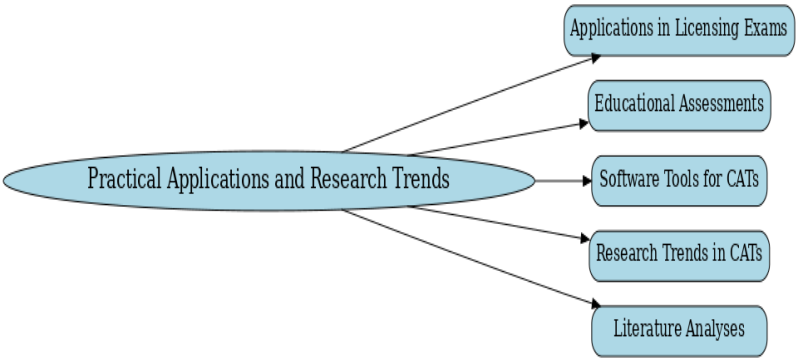

Figure 8. Classification of studies in the practical applications and research trends category based on their objectives

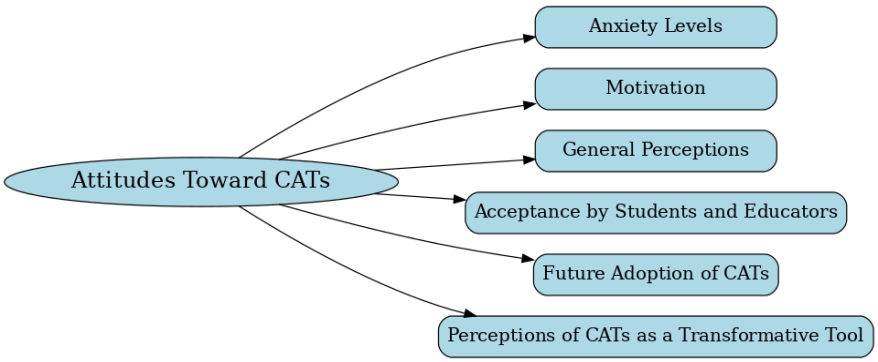

Figure 9. Classification of studies in the attitudes toward CATs category based on their objectives

Findings and Discussion on the Results of the Reviewed Studies

Studies categorized under Innovative Methods for CATs focus on new algorithms and approaches designed to improve the performance of CAT systems from various perspectives. The algorithms developed in these studies have been shown to significantly enhance measurement accuracy, reduce test duration, and ensure more balanced use of the item pool. For instance, algorithms proposed by Chao and Chen (2023), Kang et al. (2024), Pan et al. (2023), Tang et al. (2024), Xiao and Bulut (2022), and Xi et al. (2022) contributed to balanced item pool usage and controlled item exposure rates. These approaches have improved item pool efficiency and contributed to the long-term sustainability of tests. Additionally, algorithms designed to optimize test duration were found to maintain classification accuracy and increase measurement precision with shorter tests (Braeken & Paap, 2020; He & Qi, 2023; Kaplan & De La Torre, 2020; Yasuda et al., 2022). Notably, Wang et al. (2022) demonstrated that classification accuracy could be preserved in shorter tests through the use of diverse stopping rules.

Other studies focusing on measurement accuracy and test efficiency, such as those by Frey et al. (2023) and Wang et al. (2021), reported that higher levels of information could be provided throughout the test, enabling more effective evaluation of student performance. Similarly, research by Bengs et al. (2021), Kim and Yoo (2023), and Yiğit et al. (2019) revealed that comparable measurement accuracy could be achieved with shorter tests. Moreover, new methods addressing item detection and improvements in interim ability estimates proved successful in scenarios where test security was compromised (Cui, 2022; Lee & Qian, 2022). These findings suggest that new algorithms provide more efficient, reliable, and accurate measurement structures in CATs.

The findings from studies examining the applicability of CATs highlight that CATs can be applied across different domains while maintaining high measurement accuracy in less time compared to traditional tests. CATs were observed to provide reliable measurements while significantly reducing the number of items required (Dirven et al., 2021; Ebenbeck & Gebhardt, 2024; Ghio et al., 2022; Huang et al., 2022; Komarc et al., 2024; Liu et al., 2022; Tsaousis et al., 2021; Van Wijk et al., 2024). Studies in healthcare, for example, demonstrated the advantages of CATs in terms of diagnostic accuracy and speed (Adams et al., 2024). Similarly, CATs achieved more accurate classifications in reading skills assessments, effectively balancing measurement accuracy and test burden (Li et al., 2023). Furthermore, CATs were reported to significantly reduce math test anxiety, supporting their applicability in psychological assessments (Mohd-Ali et al., 2019).

However, some findings also highlighted limitations in applying CATs to specific groups. For instance, reduced measurement accuracy and longer test durations were reported for students with special education needs or abilities deviating by two standard deviations from the norm (Ebenbeck & Gebhardt, 2024). Similarly, classification performance decreased with the use of multiple cut-off points, negatively affecting accuracy (Kaya et al., 2022). Overall, these findings underscore the applicability of CATs across various domains while emphasizing certain limitations that need attention under specific conditions.

Studies examining the performance of CATs under various conditions revealed that multiple factors must be considered to enhance measurement accuracy. Longer test durations were shown to improve estimation accuracy and measurement precision, compensating for minor deviations in ability estimates. Key factors contributing to increased measurement accuracy include the use of high-quality items, which improve the efficiency of item selection algorithms in CD-CATs. Correct model selection was also found to enhance classification accuracy while shortening test durations. Adaptive item selection generally increased measurement precision, though its effect was more limited in shorter tests. The use of multi-category items and random distribution in multidimensional tests improved measurement accuracy, while multidimensional CAT designs both shortened test durations and supported measurement precision. Balanced item pools with diverse distributions improved measurement precision and reduced the number of unused items. Properly selected item exposure control methods contributed to the efficient use of item pools and maintained measurement accuracy. Content balancing positively affected test performance and measurement precision, while careful planning of item positions reduced the adverse effects of item position bias on ability estimates (Albano et al., 2019; Aşiret & Sünbül, 2024; Leroux et al., 2019; Lin et al., 2023; Mao et al., 2022; Tian & Choi, 2023).

Studies in the Practical Applications and Research Trends category examined the implementation of CAT and MST methods and reported their success in improving measurement accuracy while reducing test durations and supporting efficient item pool utilization (Ayanwale & Ndlovu, 2022; Seo & Choi, 2020; Yiğiter & Doğan, 2023). These studies also highlighted software tools and platforms that facilitate test development and implementation, providing researchers and practitioners with efficient and accurate results (Barrett et al., 2022; Lim & Wells, 2020). Additionally, issues such as potential errors in existing software, such as the catR package in R, were

addressed, and solutions were proposed (Cui, 2020). Studies emphasized the importance of interdisciplinary collaborations and integrating new technologies into the CAT field, particularly suggesting that expanding healthcare-focused research into other disciplines would significantly advance the field (Yurtçu & Güzeller, 2021).

Research in the Attitudes Toward CATs category revealed that both students and educators generally view CATs as a promising evaluation method for the future. These tests were found to increase student motivation, enrich learning experiences, and make assessment processes more effective. Additionally, the studies emphasized the critical importance of establishing the necessary infrastructure and providing training and support programs for the successful implementation of CATs (Kisielewska et al., 2024; Pramjeeth & Ramgovind, 2023).

Findings and Discussion on the Recommendations of the Reviewed Studies

Studies in the Innovative Methods for CATs category emphasize the need to develop strategies to enhance the accuracy, efficiency, and reliability of adaptive testing. Recommendations include implementing item selection algorithms to balance item pool usage, integrating behavioral data such as response time into the testing process, and establishing automated frameworks for item pool generation to maintain test security. These proposals aim to ensure that the testing process is more efficient and reliable while also promoting fairness and consistency in test outcomes. Future research is suggested to focus on improving item selection algorithms, enhancing online calibration methods, and refining procedures for controlling item exposure rates. These efforts are expected to improve test efficiency while ensuring measurement accuracy and fairness. Furthermore, examining the optimal balance between measurement precision and test efficiency through the use of response time and behavioral data is considered important. Employing machine learning methods in adaptive testing and developing new methods for detecting Differential Item Functioning (DIF) could make these tests more dynamic and adaptable.

Block designs that allow students to review or revise their responses are highlighted as having the potential to improve measurement accuracy. Specifically, Multidimensional Computerized Adaptive Tests (MCATs) supported by non-compensatory modeling methods show potential for improving measurement accuracy and efficiency through innovative approaches such as collaborative filtering, grid classification, hybrid threshold-based sequential procedures, and dynamic layering. Developing current frameworks and applying these methods to larger and more diverse datasets are also recommended. Moreover, considerations for small sample groups and content balancing are necessary for the broader application of adaptive tests in large-scale educational and assessment systems. Strategies for item selection based on response time and individual learning history, as well as models for adaptive item selection at the group level, should be more thoroughly tested to advance this field. Additionally, adapting diagnostic tree models for larger student groups and applying them to different cognitive domains is suggested. Lastly, applying information theory-based selection criteria to Cognitive Diagnosis CAT (CD-CAT) systems and comparing these criteria with other measurement indices are deemed significant steps for achieving higher precision in cognitive diagnosis processes. Implementing such innovative methods more effectively in adaptive testing could contribute to improving measurement accuracy, efficiency, and reliability in this field (Anselmi et al., 2023; Chang et al., 2019; Chao & Chen, 2023; Chen & Liu, 2023; Davis et al., 2023; Davison et al., 2023; Garcia & Thomas, 2023; Gönülateş, 2019; Gu et al., 2019; He et al., 2020; Hsu et al., 2019; Jones & Brown, 2023; Kaplan et al., 2020; Kárász et al., 2023; Lim & Choe, 2023; Lin et al., 2019; Luo et al., 2019; Smith et al., 2023; Taylor, 2023; White & Black, 2023; Williams et al., 2023; Wyse, 2023; Yuan et al., 2023; Yiğit et al., 2019).

Recommendations from articles in the Performance of CATs Under Various Conditions category focus on item quality, test length, item exposure control, and multidimensional test designs to enhance performance. In CD-CATs, using long tests and high-quality items as the number of attributes increases is recommended (Aşiret & Sünbül, 2024). Correct model selection improves performance, while tests with complex Q-matrix structures require selecting the most suitable cognitive diagnostic model for each item (Sorrel et al., 2021). Item exposure control is suggested to significantly impact the accuracy of pass/fail decisions, emphasizing the need to expand item pools (Tseng, 2021). Additionally, optimizing mixed-format items for bifactor MCATs and using A-optimality and Bayesian MAP methods to measure language skills in MCAT designs are emphasized (Özdemir & Gelbal, 2022). Strengthening item exposure control mechanisms and applying more efficient methods for balanced item pool usage are also highlighted (Leroux et al., 2019). Further research into MSTs and MCATs is recommended to examine their impact on overall test performance (Sorrel et al., 2021; Paap et al., 2019).

Recommendations include studying the effects of small item pools on measurement accuracy through item exposure control and content balancing, analyzing the impact of different test lengths on measurement precision through simulations, and evaluating the effects of response revision applications on performance in MSTs. Strengthening test security through automated item pool structuring in high-stakes assessments and optimizing item selection and stopping rules in adaptive tests to reduce standard errors and enhance ability estimation are additional suggestions (Albano et al., 2019; Cooperman et al., 2022; Mao et al., 2022; Sulak & Kelecioğlu, 2019; Tian & Choi, 2023; Yıldız et al., 2024).

Studies in the Applicability of CATs category propose several significant recommendations to increase the use of these tests. Many studies highlight how adaptive inventories provide high accuracy with fewer items, alleviating issues such as costs and respondent fatigue associated with lengthy surveys. Expanding CAT usage is recommended (Adams et al., 2024; Ayanwale & Ndlovu, 2024; Mizumoto et al., 2019; Montgomery & Rossiter, 2020; Norfarah et al., 2019; Şimşek & Tavşancıl, 2022; Van Wijk et al., 2024; Xu et al., 2023). Exploring the validity of tests across larger samples and different cultural contexts is commonly suggested (Adams et al., 2024; Ghio et al., 2022; Komarc et al., 2024; Yasuda et al., 2021). The applicability of CATs in high-stakes assessments is recommended, as these applications have the potential to provide individualized and efficient evaluations (Ayanwale & Ndlovu, 2024; Xu et al., 2023). Exploring the applicability of CD-CATs in various educational domains and improving feedback for reading skills are also emphasized (Li, Huang, & Liu, 2023). Enhancing CAT adaptability for students with special needs, using broader item pools and advanced adaptive algorithms to assess these students' performance more

accurately, is highlighted as essential (Ebenbeck & Gebhardt, 2024). Lastly, expanding item pools and refining algorithms for greater measurement accuracy are suggested as crucial for future research (Dirven et al., 2021; Fernandes et al., 2019; Şimşek & Tavşancıl, 2022).

Studies in the Practical Applications and Research Trends group provide important suggestions for the development, implementation, and investigation of CATs. The critical role of software development and online platforms in enhancing CAT efficiency is emphasized. For instance, web-based CAT platforms developed for medical licensing exams are reported to perform similarly or better than traditional tests. Expanding these systems with advanced functions, such as multi-category item models and content balancing, is recommended (Seo & Choi, 2020). Incorporating additional features and different model fit assessment methods into future versions of CAT software to reach a wider user base is also advised (Lim & Wells, 2020). Proposals to address errors encountered in existing software have been provided (Cui, 2020). These studies also suggest conducting comprehensive bibliometric analyses in various disciplines to uncover research trends in CATs (Yurtçu & Güzeller, 2021).

Research in the Attitudes Toward CATs category provides recommendations on the acceptance and adaptability of CATs by students and educators. These studies highlight that CAT usage strengthens learning motivation. Recommendations aiming to observe the long-term effects of CAT on learning motivation and knowledge levels suggest that this assessment method helps students better understand their knowledge levels. CATs are recognized as a flexible and personalized assessment method by students and educators, particularly enhancing engagement in online and remote education. Expanding CAT use in educational institutions and promoting their application across various learning environments and disciplines are emphasized as crucial suggestions (Kisielewska et al., 2024; Pramjeeth & Ramgovind, 2023).

Conclusions and Recommendations

This systematic review analyzed studies on CATs in terms of their objectives, findings, and recommendations, evaluating them under five main headings. The analysis revealed that most studies on CATs are grouped under the "Innovative Methods for CATs" category, while the fewest are in the "Attitudes Toward CATs" category. This disparity stems from the developing nature of CATs and their openness to innovation. Rapid advancements in computer hardware, software, artificial intelligence, and data processing power have contributed to the development of innovative approaches in test design and item selection algorithms. This continuous search for new solutions by researchers is aimed at improving the accuracy and efficiency of CATs.

The focus on item selection algorithms and CD-CATs in these studies can be attributed to several factors. Firstly, advances in item selection algorithms strengthen the fundamental structure of adaptive tests. Selecting the most appropriate item for an individual's ability level during the test not only improves test accuracy but also shortens the testing process and enhances the user experience. This makes improving item selection algorithms critical for the effective use of adaptive tests in fields such as education, healthcare, and psychological assessment. In large-scale test administrations, completing tests in a shorter time reduces implementation costs while increasing participant satisfaction. These benefits make item selection algorithms a focal point of research.

On the other hand, CD-CATs offer the capability to measure not only an individual's general ability level but also specific cognitive processes and knowledge domains. CD-CATs aim to identify the cognitive skills where an individual is strong or weak, rather than merely determining performance levels. This approach provides significant advantages in developing personalized learning plans in education or making detailed diagnoses in healthcare and clinical assessments. The precision of measurement provided by CD-CATs yields more accurate results in assessing multidimensional traits, meeting the need for in-depth evaluation. Consequently, studies on CD-CATs present an attractive area for researchers aiming to enhance measurement accuracy and provide more detailed evaluation processes.

In contrast, the low number of articles examining attitudes and perceptions toward CATs is noteworthy. This aspect is critical for the acceptance and success of CAT applications, as perceptions of these systems by educators and students can directly affect the implementation process. The lack of studies assessing attitudes toward CATs may hinder the collection of feedback necessary for their adoption and widespread use. The attitudes of students and educators toward CATs play a significant role in improving learning processes and evaluation experiences. Therefore, more research on attitudes and perceptions of CATs is essential for enhancing their effectiveness.

When the findings of the studies are evaluated, it is evident that newly developed algorithms improve the accuracy of CATs, shorten test processes, and provide structures that more accurately measure student performance. Studies examining the applicability of CATs have demonstrated that these tests can be implemented more quickly while maintaining high measurement accuracy compared to traditional tests in fields such as education, healthcare, and psychology. Additionally, research evaluating CAT performance under various conditions highlights the critical role of factors such as test length, item pool size, content balance, and item selection algorithms in CAT performance.

The literature review reveals the growing interest of researchers in the field of CATs and the continually expanding areas of application. This article presents an overview of the topics studied by researchers, highlighting the development of CATs and their impacts across different domains. CATs provide personalized learning experiences, better addressing the needs of test participants and making teaching and learning processes more effective. The applications of these tests have gained further importance with the spread of distance education and digital learning platforms, enabling broader accessibility.

Based on the recommendations, developing new methods in the field of CATs emerges as a critical requirement for improving the accuracy, efficiency, and reliability of these tests. Utilizing the new software and hardware capabilities offered by technology to develop innovative methods will enhance the effectiveness of CAT processes. Expanding the use of CATs across different disciplines and learning environments will contribute to learning processes by offering assessment opportunities tailored to individual needs.

Author Contributions

All authors have equally contributed to all processes of the article. All authors have read and approved the final version of the study.

Ethical Declaration

The authors declare that their work is not subject to ethics committee approval and that the rules set by the Committee on Publication Ethics (COPE) were followed throughout the study.

Conflict of Interest

The authors declare that there is no conflict of interest with any institution or person within the scope of the study.

References

Adams, Z. W., Hulvershorn, L. A., Smoker, M. P., Marriott, B. R., Aalsma, M. C., & Gibbons, R. D. (2024). Initial validation of a computerized adaptive test for substance use disorder identification in adolescents. *Substance Use & Misuse*, *59*(6), 867–873. https://doi.org/10.1080/10826084.2024.2305801

Albano, A. D., Cai, L., Lease, E. M., & McConnell, S. R. (2019). Computerized adaptive testing in early education: Exploring the impact of item position effects on ability estimation. *Journal of Educational Measurement*, *56*(2), 437–451. https://doi.org/10.1111/jedm.12215

Anselmi, P., Robusto, E., & Cristante, F. (2023). Enhancing computerized adaptive testing with batteries of unidimensional tests. *Applied Psychological Measurement*, *47*(3), 167–182. https://doi.org/10.1177/01466216231165301

Aşiret, 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. https://doi.org/10.21031/epod.1456094

Ayanwale, M. A., & Ndlovu, M. (2022). Transition from computer-based testing of national benchmark tests to adaptive testing: Robust application of fourth industrial revolution tools. *Cypriot Journal of Educational Sciences*, *17*(9), 3327–3343. https://doi.org/10.18844/cjes.v17i9.7124

Ayanwale, M. A., & Ndlovu, M. (2024). The feasibility of computerized adaptive testing of the national benchmark test: A simulation study. *Journal of Pedagogical Research*, *8*(2), 95–112. https://doi.org/10.33902/JPR.202425210

Barrett, M. D., Jiang, B., & Feagler, B. E. (2022). A smart authoring system for designing, configuring, and deploying adaptive assessments at scale. *International Journal of Artificial Intelligence in Education*, *32*(1), 28–47. https://doi.org/10.1007/s40593-021-00258-y

Bengs, D., Kroehne, U., & Brefeld, U. (2021). Simultaneous constrained adaptive item selection for group-based testing. *Journal of Educational Measurement*, *58*(2), 236–261. https://doi.org/10.1111/jedm.12285

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*(4), 275–285. https://doi.org/10.1111/j.1745-3984.1988.tb00308.x

Braeken, J., & Paap, M. C. (2020). Making fixed-precision between-item multidimensional computerized adaptive tests even shorter by reducing the asymmetry between selection and stopping rules. *Applied Psychological Measurement*, *44*(7–8), 531–547. https://doi.org/10.1177/0146621620932666

Chang, H. H., & Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *Annals of Statistics*, *37*(3), 1466–1488. https://doi.org/10.1214/08-AOS614

Chang, Y. P., Chiu, C. Y., & Tsai, R. C. (2019). Nonparametric CAT for CD in educational settings with small samples. *Applied Psychological Measurement*, *43*(7), 543–561. https://doi.org/10.1177/0146621618813113

Chao, H. Y., & Chen, J. H. (2023). Controlling the minimum item exposure rate in computerized adaptive testing: A two-stage Sympson–Hetter procedure. *Applied Psychological Measurement*, *47*(7–8), 460–477. https://doi.org/10.1177/01466216231209756

Chen, C. W., & Liu, C. W. (2023). Online Parameter Estimation for Student Evaluation of Teaching. *Applied Psychological Measurement*, *47*(4), 291-311. https://doi.org/10.1177/01466216231165314

Chen, C. W., Wang, W. C., Chiu, M. M., & Ro, S. (2020). Item selection and exposure control methods for computerized adaptive testing with multidimensional ranking items. *Journal of Educational Measurement*, *57*(2), 343–369. https://doi.org/10.1111/jedm.12252

Chen, J. H., & Chao, H. Y. (2024). Utilizing real-time Test Data to solve attenuation paradox in computerized adaptive testing to enhance optimal design. *Journal of Educational and Behavioral Statistics*, *49*(4), 630-657. https://doi.org/10.3102/10769986231197666

Chen, J. H., Chao, H. Y., & Chen, S. Y. (2020). A dynamic stratification method for improving trait estimation in computerized adaptive testing under item exposure control. *Applied Psychological Measurement*, *44*(3), 182–196. https://doi.org/10.1177/0146621619843820

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, *40*(2), 129–145. https://doi.org/10.1111/j.1745-3984.2003.tb01100.x

Cooperman, A. W., Weiss, D. J., & Wang, C. (2022). Robustness of adaptive measurement of change to item parameter estimation error. *Educational and Psychological Measurement*, *82*(4), 643–677. https://doi.org/10.1177/00131644211033902

Cui, Z. (2020). A seed usage issue on using catR for simulation and the solution. *Applied Psychological Measurement*, *44*(5), 409–412. https://doi.org/10.1177/0146621620920934

Cui, Z. (2022). On measuring adaptivity of an adaptive test. *Measurement: Interdisciplinary Research and Perspectives*, *20*(1), 21–33. https://doi.org/10.1080/15366367.2021.1922232

Davison, M. L., Weiss, D. J., DeWeese, J. N., Ersan, O., Biancarosa, G., & Kennedy, P. C. (2023). A diagnostic tree model for adaptive assessment of complex cognitive processes using multidimensional response options. *Journal of Educational and Behavioral Statistics*, *48*(6), 914–941. https://doi.org/10.3102/10769986231158301

Dirven, L., Petersen, M. A., Aaronson, N. K., Chie, W. C., Conroy, T., Costantini, A., Hammerlid, E., Velikova, G., Verdonck-de Leeuw, I. M., Young, T., & Groenvold, M. (2021). Development and psychometric evaluation of an

item bank for computerized adaptive testing of the EORTC insomnia dimension in cancer patients (EORTC CAT-SL). *Applied Research in Quality of Life, 16*, 827–844. https://doi.org/10.1007/s11482-019-09799-w

Ebenbeck, N., & Gebhardt, M. (2022). Simulating computerized adaptive testing in special education based on inclusive progress monitoring data. *Frontiers in Education, 7*, Article 945733. https://doi.org/10.3389/feduc.2022.945733

Ebenbeck, N., & Gebhardt, M. (2024). Differential performance of computerized adaptive testing in students with and without disabilities – A simulation study. *Journal of Special Education Technology, 39*(4), 481–490. https://doi.org/10.1177/01626434241232117

Fernandes, S., Fond, G., Zendjidjian, X., Michel, P., Baumstarck, K., Lancon, C., Berna, F., Schurhoff, F., Aouizerate, B., Henry, C., Etain, B., Samalin, L., Leboyer, M., Llorca, P. M., Coldefy, M., Auquier, P., & Boyer, L. (2019). The Patient-Reported Experience Measure for Improving Quality of Care in Mental Health (PREMIUM) project in France: Study protocol for the development and implementation strategy. *Patient Preference and Adherence, 13*, 165–177. https://doi.org/10.2147/PPA.S172100

Frey, A., König, C., & Fink, A. (2023). A highly adaptive testing design for PISA. *Journal of Educational Measurement.* https://doi.org/10.1111/jedm.12382

Gao, X., Wang, D., Cai, Y., & Tu, D. (2020). Cognitive diagnostic computerized adaptive testing for polytomously scored items. *Journal of Classification, 37*, 709–729. https://doi.org/10.1007/s00357-019-09357-x

Ghio, F. B., Bruzzone, M., Rojas-Torres, L., & Cupani, M. (2022). Preliminary development of an item bank and an adaptive test in mathematical knowledge for university students. *European Journal of Science and Mathematics Education, 10*(3), 352–365. https://doi.org/10.30935/scimath/11968

Gönülateş, E. (2019). Quality of Item Pool (QIP) index: A novel approach to evaluating CAT item pool adequacy. *Educational and Psychological Measurement, 79*(6), 1133–1155. https://doi.org/10.1177/0013164419842215

Gu, L., Ling, G., & Qu, Y. (2019). A modified a-stratified method for computerized adaptive testing. *ETS Research Report Series, 2019*(1), 1–27. https://doi.org/10.1002/ets2.12246

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Han, K. C. T. (2020). Framework for developing multistage testing with intersectional routing for short-length tests. *Applied Psychological Measurement, 44*(2), 87–102. https://doi.org/10.1177/0146621619837226

He, Y., & Qi, Y. (2023). Using response time in multidimensional computerized adaptive testing. *Journal of Educational Measurement, 60*(4), 697–738. https://doi.org/10.1111/jedm.12373

He, Y., Chen, P., & Li, Y. (2020). New efficient and practicable adaptive designs for calibrating items online. *Applied Psychological Measurement, 44*(1), 3–16. https://doi.org/10.1177/0146621618824854

Hsu, C. L., & Wang, W. C. (2019). Multidimensional computerized adaptive testing using non-compensatory item response theory models. *Applied Psychological Measurement, 43*(6), 464–480. https://doi.org/10.1177/0146621618800280

Hsu, C. L., & Wang, W. C. (2022). Reducing the misclassification costs of cognitive diagnosis computerized adaptive testing: Item selection with minimum expected risk. *Applied Psychological Measurement, 46*(3), 185–199. https://doi.org/10.1177/01466216211066610

Huang, H. T. D., Hung, S. T. A., Chao, H. Y., Chen, J. H., Lin, T. P., & Shih, C. L. (2022). Developing and validating a computerized adaptive testing system for measuring the English proficiency of Taiwanese EFL university students. *Language Assessment Quarterly, 19*(2), 162–188. https://doi.org/10.1080/15434303.2021.1984490

Istiyono, E., Dwandaru, W. S. B., Setiawan, R., & Megawati, I. (2020). Developing of computerized adaptive testing to measure physics higher order thinking skills of senior high school students and its feasibility of use. *European Journal of Educational Research, 9*(1), 91–101. https://doi.org/10.12973/eu-jer.9.1.91

Jodoin, M. G., Zenisky, A. L., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education, 19*(3), 203–220. https://doi.org/10.1207/s15324818ame1903_3

Kang, H. A., Arbet, G., Betts, J., & Muntean, W. (2024). Location-matching adaptive testing for polytomous technology-enhanced items. *Applied Psychological Measurement, 48*(1–2), 57–76. https://doi.org/10.1177/01466216241227548

Kang, H. A., Zheng, Y., & Chang, H. H. (2020). Online calibration of a joint model of item responses and response times in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 45*(2), 175–208. https://doi.org/10.3102/1076998619879040

Kaplan, M., & De La Torre, J. (2020). A blocked-CAT procedure for CD-CAT. *Applied Psychological Measurement, 44*(1), 49–64. https://doi.org/10.1177/0146621619835500

Kaya, E., O'Grady, S., & Kalender, İ. (2022). IRT-based classification analysis of an English language reading proficiency subtest. *Language Testing, 39*(4), 541–566. https://doi.org/10.1177/02655322211068847

Kern, J. L., & Choe, E. (2021). Using a response time–based expected a posteriori estimator to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement, 45*(5), 361–385. https://doi.org/10.1177/01466216211014601

Kim, R. Y., & Yoo, Y. J. (2023). Cognitive diagnostic multistage testing by partitioning hierarchically structured attributes. *Journal of Educational Measurement, 60*(1), 126–147. https://doi.org/10.1111/jedm.12339

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359–375. https://doi.org/10.1207/s15324818ame0204_6

Kisielewska, J., Millin, P., Rice, N., Pego, J. M., Burr, S., Nowakowski, M., & Gale, T. (2024). Medical students' perceptions of a novel international adaptive progress test. *Education and Information Technologies, 29*(9), 11323–11338. https://doi.org/10.1007/s10639-023-12269-4

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, 1–18.* https://doi.org/10.1080/19419899.2024.2332630

Lee, C., & Qian, H. (2022). Hybrid threshold-based sequential procedures for detecting compromised items in a computerized adaptive testing licensure exam. *Educational and Psychological Measurement, 82*(4), 782–810. https://doi.org/10.1177/00131644211023868

Leroux, A. J., Waid-Ebbs, J. K., Wen, P. S., Helmer, D. A., Graham, D. P., O'Connor, M. K., & Ray, K. (2019). An investigation of exposure control methods with variable-length CAT using the partial credit model. *Applied Psychological Measurement, 43*(8), 624–638. https://doi.org/10.1177/0146621618824856

Li, Y., Huang, C., & Liu, J. (2023). Diagnosing primary students' reading progression: Is cognitive diagnostic computerized adaptive testing the way forward? *Journal of Educational and Behavioral Statistics, 48*(6), 842–865. https://doi.org/10.3102/10769986231160668

Lim, H., & Choe, E. M. (2023). Detecting differential item functioning in CAT using IRT residual DIF approach. *Journal of Educational Measurement, 60*(4), 626–650. https://doi.org/10.1111/jedm.12366

Lim, H., & Han, K. C. T. (2024). An automated item pool assembly framework for maximizing item utilization for CAT. *Educational Measurement: Issues and Practice, 43*(1), 39–51. https://doi.org/10.1111/emip.12589

Lim, H., & Wells, C. S. (2020). irtplay: An R package for online item calibration, scoring, evaluation of model fit, and useful functions for unidimensional IRT. *Applied Psychological Measurement, 44*(7–8), 563–565. https://doi.org/10.1177/0146621620921247

Lin, C. J., & Chang, H. H. (2019). Item selection criteria with practical constraints in cognitive diagnostic computerized adaptive testing. *Educational and Psychological Measurement, 79*(2), 335–357. https://doi.org/10.1177/0013164418790634

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

Linden, W. J., & Glas, G. A. W. (2002). Computerized adaptive testing: Theory and practice. *Kluwer Academic Publishers.*

Liu, K., Zhang, L., Tu, D., & Cai, Y. (2022). Developing an item bank of computerized adaptive testing for eating disorders in Chinese university students. *SAGE Open, 12*(4), 1–13. https://doi.org/10.1177/21582440221141273

Luo, H., Wang, D., Guo, Z., Cai, Y., & Tu, D. (2022). Combining cognitive diagnostic computerized adaptive testing with multidimensional item response theory. *Applied Psychological Measurement, 46*(4), 288–302. https://doi.org/10.1177/01466216221084214

Luo, X., & Wang, X. (2019). Dynamic multistage testing: A highly efficient and regulated adaptive testing method. *International Journal of Testing, 19*(3), 227–247. https://doi.org/10.1080/15305058.2019.1621871

Mao, X., Zhang, J., & Xin, T. (2022). The optimal design of bifactor multidimensional computerized adaptive testing with mixed-format items. *Applied Psychological Measurement, 46*(7), 605–621. https://doi.org/10.1177/01466216221108382

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449–458. https://doi.org/10.1037/0033-2909.114.3.449

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23*(3), 187–194. https://doi.org/10.1177/01466219922031310

Mizumoto, A., Sasao, Y., & Webb, S. A. (2019). Developing and evaluating a computerized adaptive testing version of the Word Part Levels Test. *Language Testing, 36*(1), 101–123. https://doi.org/10.1177/0265532217725776

Mohd-Ali, S., Norfarah, N., Ilya-Syazwani, J. I., & Mohd-Erfy, I. (2019). The effect of computerized adaptive testing on reducing anxiety towards math test for polytechnic students. *Journal of Technical Education and Training, 11*(4), 27–35. https://doi.org/10.30880/jtet.2019.11.04.004

Montgomery, J. M., & Rossiter, E. L. (2020). So many questions, so little time: Integrating adaptive inventories into public opinion research. *Journal of Survey Statistics and Methodology, 8*(4), 667–690. https://doi.org/10.1093/jssam/smz027

Özdemir, B., & Gelbal, S. (2022). Measuring language ability of students with compensatory multidimensional CAT: A post-hoc simulation study. *Education and Information Technologies, 27*(5), 6273–6294. https://doi.org/10.1007/s10639-021-10853-0

Öztürk, N. B., & Şahin, M. G. (2019). Effects of item pool characteristics on ability estimate and item pool utilization: A simulation study. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 34*(2), 473–486. https://doi.org/10.16986/HUJE.2018042418

Paap, M. C., Born, S., & Braeken, J. (2019). Measurement efficiency for fixed-precision multidimensional computerized adaptive tests: Comparing health measurement and educational testing using example banks. *Applied Psychological Measurement, 43*(1), 68–83. https://doi.org/10.1177/0146621618765719

Page, M. J., McKenzie, J. E., 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., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Pan, Y., Livne, O., Wollack, J. A., & Sinharay, S. (2023). Item selection algorithm based on collaborative filtering for item exposure control. *Educational Measurement: Issues and Practice, 42*(4), 6–18. https://doi.org/10.1111/emip.12578

Pramjeeth, S., & Ramgovind, P. (2023). Students' and lecturers' perceptions of computerized adaptive testing as the future of assessing students. *Journal of Education (University of KwaZulu-Natal), 93*, 120–146. http://dx.doi.org/10.17159/2520-9868/i93a06

Pranckutė, R. (2021). Web of science (WoS) and scopus: The titans of bibliographic information in today's academic world. *Publications, 9*(1), 12. https://doi.org/10.3390/publications9010012

Qiu, X. L., De La Torre, J., Ro, S., & Wang, W. C. (2022). Computerized adaptive testing for ipsative tests with multidimensional pairwise-comparison items: Algorithm development and applications. *Applied Psychological*

Measurement, 46(4), 255–272. https://doi.org/10.1177/01466216221084209

Raborn, A., & Sarı, H. (2021). Mixed adaptive multistage testing: A new approach. *Journal of Measurement and Evaluation in Education and Psychology, 12*(4), 358–373. https://doi.org/10.21031/epod.871014

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice, 8*(3), 11–15. https://doi.org/10.1111/j.1745-3992.1989.tb00326.x

Segall, D. O. (2005). Computerized adaptive testing. In Kempf-Leonard (Ed.), *The Encyclopedia of Social Measurement* (pp. 429–438). Academic Press.

Seo, D. G., & Choi, J. (2020). Introduction to the LIVECAT web-based computerized adaptive testing platform. *Journal of Educational Evaluation for Health Professions, 17*, 1–7. https://doi.org/10.3352/jeehp.2020.17.27

Sorrel, M. A., Abad, F. J., & Nájera, P. (2021). Improving accuracy and usage by correctly selecting: The effects of model selection in cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 45*(2), 112–129. https://doi.org/10.1177/0146621620977682

Sulak, S., & Kelecioğlu, H. (2019). Investigation of item selection methods according to test termination rules in CAT applications. *Journal of Measurement and Evaluation in Education and Psychology, 10*(3), 315–326. https://doi.org/10.21031/epod.530528

Sun, X., Andersson, B., & Xin, T. (2021). A new method to balance measurement accuracy and attribute coverage in cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement, 45*(7–8), 463–476. https://doi.org/10.1177/01466216211040489

Şenel, S. (2021). *Bilgisayar ortamında bireye uyarlanmış testler*. Hacettepe Üniversitesi Yayınları.

Şimşek, A. S., & Tavşancıl, 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. https://doi.org/10.21031/epod.1148313

Tan, Q., Cai, Y., Luo, F., & Tu, D. (2023). Development of a high-accuracy and effective online calibration method in CD-CAT based on Gini index. *Journal of Educational and Behavioral Statistics, 48*(1), 103–141. https://doi.org/10.3102/10769986221126741

Tang, X., Zheng, Y., Wu, T., Hau, K. T., & Chang, H. H. (2024). Utilizing response time for item selection in on-the-fly multistage adaptive testing for PISA assessment. *Journal of Educational Measurement.* https://doi.org/10.1111/jedm.12403

Tian, C., & Choi, J. (2023). The impact of item model parameter variations on person parameter estimation in computerized adaptive testing with automatically generated items. *Applied Psychological Measurement, 47*(4), 275–290. https://doi.org/10.1177/01466216231165313

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. https://doi.org/10.1177/07342829211027753

Tseng, W. T. (2021). The effects of item exposure control on measurement precision of vocabulary size estimates in computerized adaptive testing. *English Teaching & Learning, 45*(2), 217–236. https://doi.org/10.1007/s42321-020-00068-w

Van der Linden, W. J., & Glas, C. A. W. (2002). Computerized adaptive testing: Theory and practice. Kluwer Academic.

Van Wijk, E. V., Donkers, J., De Laat, P. C. J., Meiboom, A. A., Jacobs, B., Ravesloot, J. H., Tio, R. A., Van Der Vleuten, C. P. M., Langers, A. M. J., & Bremers, A. J. A. (2024). Computer adaptive vs. non-adaptive medical progress testing: Feasibility, test performance, and student experiences. *Perspectives on Medical Education, 13*(1), 406–416. https://doi.org/10.5334/pme.1345

Wang, C. (2021). On interim cognitive diagnostic computerized adaptive testing in learning context. *Applied Psychological Measurement, 45*(4), 235–252. https://doi.org/10.1177/0146621621990755

Wang, C., & Zhu, R. (2024). Detecting uniform differential item functioning for continuous response computerized adaptive testing. *Applied Psychological Measurement, 48*(1–2), 18–37. https://doi.org/10.1177/01466216241227544

Wang, S., Xiao, H., & Cohen, A. (2021). Adaptive weight estimation of latent ability: Application to computerized adaptive testing with response revision. *Journal of Educational and Behavioral Statistics, 46*(5), 560–591. https://doi.org/10.3102/1076998620972800

Wang, W., Song, L., Wang, T., Gao, P., & Xiong, J. (2020). A note on the relationship of the Shannon entropy procedure and the Jensen–Shannon divergence in cognitive diagnostic computerized adaptive testing. *SAGE Open, 10*(1), 2158244019899046. https://doi.org/10.1177/2158244019899046

Wang, Z., Wang, C., & Weiss, D. J. (2022). Termination criteria for grid multiclassification adaptive testing with multidimensional polytomous items. *Applied Psychological Measurement, 46*(7), 551–570. https://doi.org/10.1177/01466216221108383

Weiss, D. J., & Şahin, A. (2024). *Computerized adaptive testing: From concept to implementation.* Guilford Publications.

Wyse, A. E. (2021). How days between tests impacts alternate forms reliability in computerized adaptive tests. *Educational and Psychological Measurement, 81*(4), 644–667. https://doi.org/10.1177/0013164420979656

Wyse, A. E. (2023). Two statistics for measuring the score comparability of computerized adaptive tests. *Applied Psychological Measurement, 47*(7-8), 513-525. https://doi.org/10.1177/01466216231209749

Xi, C., Tu, D., & Cai, Y. (2022). Dual-objective item selection methods in computerized adaptive test using the higher-order cognitive diagnostic models. *Applied Psychological Measurement, 46*(5), 422-438. https://doi.org/10.1177/01466216221089342

Xiao, J., & Bulut, O. (2022). Item selection with collaborative filtering in on-the-fly multistage adaptive testing. *Applied Psychological Measurement, 46*(8), 690–704. https://doi.org/10.1177/01466216221124089

Xu, L., Jiang, Z., Han, Y., Liang, H., & Ouyang, J. (2023). Developing computerized adaptive testing for a national health professionals exam: An attempt from psychometric simulations. *Perspectives on Medical Education, 12*(1), 462. https://doi.org/10.5334/pme.855

Yang, J., Chang, H. H., Tao, J., & Shi, N. (2020). Stratified item selection methods in cognitive diagnosis

computerized adaptive testing. *Applied Psychological Measurement, 44*(5), 346–361. https://doi.org/10.1177/0146621619893783

Yang, L., & Reckase, M. D. (2020). The optimal item pool design in multistage computerized adaptive tests with the p-optimality method. *Educational and Psychological Measurement*, *80*(5), 955-974.

Yasuda, J. I., Hull, M. M., & Mae, N. (2022). Improving test security and efficiency of computerized adaptive testing for the Force Concept Inventory. *Physical Review Physics Education Research, 18*(1), 010112. https://doi.org/10.1103/PhysRevPhysEducRes.18.010112

Yasuda, J. I., Mae, N., Hull, M. M., & Taniguchi, M. A. (2021). Optimizing the length of computerized adaptive testing for the Force Concept Inventory. *Physical Review Physics Education Research, 17*(1), 010115. https://doi.org/10.1103/PhysRevPhysEducRes.17.010115

Yıldız, H., Demir, C., Ülkü, S., Giray, G., & Kelecioğlu, H. (2024). Investigation of measurement precision and test length in computerized adaptive tests under different conditions. *Journal of Measurement and Evaluation in Education and Psychology-EPOD, 15*(1), 5–17. https://doi.org/10.21031/epod.1068572

Yiğit, H. D., Sorrel, M. A., & De La Torre, J. (2019). Computerized adaptive testing for cognitively based multiple-choice data. *Applied Psychological Measurement, 43*(5), 388–401. https://doi.org/10.1177/0146621618798665

Yiğiter, M. S., & Doğan, N. (2023). Computerized multistage testing: Principles, designs and practices with R. *Measurement: Interdisciplinary Research and Perspectives, 21*(4), 254–277. https://doi.org/10.1080/15366367.2022.2158017

Yiğiter, M. S., & Doğan, N. (2024). Comparison of different computerized adaptive testing approaches with shadow test under different test lengths and ability estimation method conditions. *Journal of Measurement and Evaluation in Education and Psychology-EPOD, 14*(4), 396–412. https://doi.org/10.21031/epod.1202599

Yuan, L., Huang, Y., Li, S., & Chen, P. (2023). Online calibration in multidimensional computerized adaptive testing with polytomously scored items. *Journal of Educational Measurement, 60*(3), 476–500. https://doi.org/10.1111/jedm.12353

Yuhana, U. L., Yuniarno, E. M., Rahayu, W., & Pardede, E. (2024). A Context-based question selection model to support the adaptive assessment of learning: A study of online learning assessment in elementary schools in Indonesia. *Education and Information Technologies*, *29*(8), 9517-9540. https://doi.org/10.1007/s10639-023-12184-8

Yurtçu, M., & Güzeller, C. (2021). Bibliometric analysis of articles on computerized adaptive testing. *Participatory Educational Research, 8*(4), 426–438. https://doi.org/10.17275/per.21.98.8.4