Iranian Journal of Psychoeducational Assessment
University of Zanjan

Spring & Summer 2025. Vol 1. Issue 1

Research Paper

Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT): Conceptual Framework and Steps for Test Development and Validation

Samira Tavanaei Yousefian[1*], Asghar Minaei[2], Mohammad Asgari[3], Shabnam Shadroo [4], Ali Mashhadi[5]

1. PhD Candidate in Educational Measurement and Assessment, Department of Educational and Psychological Measurement, Faculty of Psychology and Education, Allameh Tabataba'i University, Tehran, Iran
2. Associate Professor, Department of Educational and Psychological Measurement, Faculty of Psychology and Education, Allameh Tabataba'i University, Tehran, Iran
3. Professor, Department of Educational and Psychological Measurement, Faculty of Psychology and Education, Allameh Tabataba'i University, Tehran, Iran
4. Assistant Professor, Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran
5. Professor, Department of Psychology, Ferdowsi University of Mashhad, Mashhad, Iran

Article info:
Received: 18.06.2025
Revised: 29.06.2025
Accepted: 10.07.2025

Keywords:
cognitive diagnostic Computerized Adaptive Testing (CD-CAT), Computerized Adaptive Testing (CAT), Cognitive Diagnostic Models (CDMs)

CC BY

Publisher: University of Zanjan

Abstract

The present study explores the process of test construction using the Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT) approach, a novel method in educational assessment. This approach integrates methods from Computerized Adaptive Testing (CAT) and Cognitive Diagnostic Models (CDMs) to identify individuals' abilities and weaknesses with greater accuracy and in a more personalized manner. Initially, the theoretical foundations and key concepts related to CD-CAT are discussed. CD-CAT can provide a more precise assessment of candidates' abilities and skills. Instead of offering a general score, this approach provides a deeper understanding of a student's academic status and strengths by analyzing specific skill strengths and weaknesses. The article details the various stages of CD-CAT test construction, including item bank design, item selection algorithms, and test termination rules. Furthermore, the advantages of this approach over traditional tests, such as increased accuracy and reliability of assessments and improved efficiency of educational processes, are discussed. Research indicates that CD-CAT can optimize the evaluation process by selecting items tailored to the candidate's ability level based on their previous responses. The findings also suggest that CD-CAT can significantly enhance the quality of assessments and the accuracy of diagnoses. Finally, suggestions for future research in the development and improvement of CD-CAT and its practical applications in various educational and assessment fields are provided.

Use your device to scan and read the article online

Citation: Tavanaei Yousefian, S., Minaei, A., Asgari, M., Shadroo, S., Mashhadi, A. (2025). Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT): Conceptual framework and steps for test development and validation. *Iranian Journal of Psychoeducational Assessment*, 1 (1), 171-191. ttps://doi.org/10.30470/ijpa.2025.726664

***Corresponding Author:** Samira Tavanaei Yousefian
Address: Department of Educational and Psychological Measurement, Faculty of Psychology and Education, Allameh Tabataba'i University, Tehran, Iran
Email: tavanaei.yousefian@gmail.com

Extended Abstract

Introduction

In recent decades, the limitations of traditional paper-based assessments have driven a global shift toward more personalized and sophisticated methods of educational evaluation. Among these, Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT) has emerged as a cutting-edge solution that fuses two major psychometric paradigms: Computerized Adaptive Testing (CAT) and Cognitive Diagnostic Models (CDMs) (Andrade Medeiros,2022; Rudner,1998; Cheng, 2009). Unlike conventional tests that offer only a single composite score, CD-CAT provides detailed diagnostic profiles of learners by assessing their mastery of multiple latent cognitive attributes. This enables educators to make precise, data-driven instructional decisions and tailor interventions to individual learners (Weiss & Sahin, 2024; Tatsuoka & Ferguson, 2003).

Methods

This study utilizes a descriptive-analytical review to examine the theoretical framework, functional design, and operational stages of CD-CAT systems. It outlines five key phases: (1) constructing a Q-matrix to map test items to cognitive attributes, which is vital for diagnostic accuracy; (2) defining rules for test initiation, including how initial items are selected based on prior assumptions; (3) implementing adaptive item selection strategies such as Shannon Entropy, KL Information, GDI, and PWKL to maximize information gained from each response; (4) estimating the test-taker's mastery profile progressively using methods like MLE, MAP, and EAP; and (5) applying stopping rules to determine when sufficient data has been collected to end the assessment efficiently. This structured approach enhances the precision and adaptability of cognitive diagnostic testing. Additionally, this review evaluates the practical applications of CD-CAT through tools such as GDINA and Concerto, which support both simulation and real-world implementation. These platforms offer flexibility and robust algorithmic infrastructure, making them suitable for researchers and practitioners engaged in CD-CAT.

Results

The findings of this review highlight the considerable advantages of Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT) in enhancing both diagnostic precision and testing efficiency. By generating multidimensional skill profiles, CD-CAT allows for a more nuanced understanding of learners' cognitive strengths and weaknesses, moving beyond traditional score-based assessments.

From a practical perspective, educators and psychometricians have found that advanced item selection algorithms such as Jensen-Shannon Divergence (JSD) and Modified Posterior-Weighted Kullback-Leibler (MPWKL) have demonstrated superior performance in recovering accurate mastery patterns while simultaneously reducing test length (Aşiret & Sünbül, 2024). These algorithms optimize the balance between informational gain and test burden, making the assessment process both shorter and more precise (Zheng & Wang, 2017). A critical factor influencing the validity of CD-CAT results is the design quality of the Q-matrix, which maps each test item to its corresponding latent attributes. Inaccurate mappings can distort diagnostic interpretations and lead to invalid educational inferences (Rupp & Templin, 2008; De La Torre, 2009). Therefore, meticulous construction and validation of the Q-matrix are essential for ensuring diagnostic reliability. To support the complex data analysis and implementation requirements of CD-CAT, software platforms like GDINA and Concerto have proven particularly effective. These tools facilitate the modeling, estimation, and interpretation of cognitive skill profiles, providing researchers and practitioners with accessible yet powerful environments for operationalizing CD-CAT in real-world settings (Ma & de la Torre, 2020).

Conclusion

Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT) represents a significant advancement in educational assessment by combining the diagnostic depth of Cognitive Diagnostic Models (CDMs) with the adaptability and efficiency of Computerized Adaptive Testing

(CAT). Unlike traditional unidimensional approaches that reduce learner performance to a single score, CD-CAT provides a comprehensive mastery profile across multiple cognitive attributes, allowing for a more accurate, learner-centered evaluation (Cheng, 2009; Yu et al, 2019). This integration enables educators to make data-informed instructional decisions tailored to the specific needs of individual students. By pinpointing which skills are mastered and which require further development, CD-CAT supports personalized learning pathways and more targeted interventions (Tatsuoka & Ferguson, 2003).

Despite these strengths, several challenges remain that warrant further investigation. Chief among them is the validation of the Q-matrix, which serves as the backbone of the diagnostic process. Inaccuracies in mapping items to latent attributes can compromise the validity of results and misinform educational decisions (Rupp & Templin, 2008). Additionally, selecting the most suitable item selection strategy (ISS) remains a complex task, as each algorithm presents trade-offs in terms of efficiency, accuracy, and computational demand (Li et al, 2023). Translating CD-CAT from theory into practice requires addressing logistical, technical, and pedagogical challenges that extend beyond theoretical modeling. Furthermore, a notable gap exists in the application of CD-CAT within the Iranian educational context, underscoring the importance of conducting localized research to ensure cultural and curricular alignment (Ravand & Baghaei, 2020).

Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT) represents a new generation of assessment systems that are not only psychometrically rigorous but also highly responsive to individual learner needs. By combining the precision of cognitive diagnostic models with the adaptability of computerized testing, CD-CAT provides a nuanced understanding of students' cognitive strengths and weaknesses. This insight enables educators to design targeted instructional strategies and personalized interventions that are both effective and evidence-based (Cheng, 2009; Yu et al, 2019). The learner-centered nature of CD-CAT marks a shift from traditional score-focused evaluation to a more holistic and diagnostic approach. Its integration into educational practice can significantly enhance teaching quality and learning outcomes by offering real-time, skill-specific feedback to students and instructors alike.

To successfully scale CD-CAT, further research and practical improvements are essential especially in refining Q-matrices, stopping rules, and software design. Its application in diverse contexts like Iran also demands cultural adaptation and localized testing (Ravand & Baghaei, 2020). Achieving full impact calls for ongoing interdisciplinary efforts and strong institutional backing.

Ethical Considerations

This study is based on a literature review and does not involve any human participants. Nevertheless, ethical considerations, including integrity and proper citation of sources, have been duly observed.

Authors' Contributions

This article is derived from a doctoral dissertation. The first author is a Ph.D. candidate, the second and third authors served as academic supervisors, and the fourth and fifth authors contributed as scientific research advisors to this dissertation.

Funding

The costs of conducting this research and preparing the manuscript were fully covered by the authors. This study did not receive any external financial support.

Conflicts of Interest

The authors declare that there is no conflict of interest.

دو فصلنامه علمی سنجش روانی تربیتی
دانشگاه زنجان

مقاله پژوهشی

سنجش انطباقی کامپیوتری تشخیصی شناختی (CD-CAT): چارچوب و گام‌های ساخت و رواسازی آزمون‌های آموزشی

سمیرا توانایی یوسفیان[۱]، اصغر مینایی[۲*]، محمد عسگری[۳]، شبنم شادرو[۴]، علی مشهدی[۵]

۱. دانشجوی دکتری رشته سنجش و اندازه گیری، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبائی، تهران، ایران
۲. دانشیار گروه سنجش و اندازه‌گیری، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبائی، تهران، ایران
۳. استاد گروه سنجش و اندازه‌گیری، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبائی، تهران، ایران
۴. استادیار گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی، واحد مشهد، ایران
۵. استاد گروه روان‌شناسی، دانشکده علوم تربیتی و روان‌شناسی، دانشگاه فردوسی مشهد، مشهد، ایران

اطلاعات مقاله:
تاریخ دریافت: ۱۴۰۴/۰۳/۲۸
تاریخ داوری: ۱۴۰۴/۰۴/۰۹
تاریخ پذیرش: ۱۴۰۴/۰۴/۲۰

واژه‌های کلیدی:
سنجش انطباقی کامپیوتری تشخیصی شناختی (CD-CAT)، سنجش انطباقی کامپیوتری (CAT)، مدل‌های تشخیصی شناختی (CDMs)، آزمون‌سازی

CC BY

ناشر: دانشگاه زنجان

چکیده

پژوهش حاضر به فرایند ساخت آزمون با رویکرد سنجش انطباقی کامپیوتری تشخیصی شناختی (CD-CAT) می‌پردازد که یک روش نوین در ارزیابی تحصیلی است. این رویکرد، ترکیبی از روش‌های سنجش انطباقی (CAT) و مدل‌های تشخیصی شناختی (CDMs) را به کار می‌گیرد تا بتواند به طور دقیق‌تر و شخصی‌سازی‌شده‌تری توانایی‌ها و نقاط ضعف افراد را شناسایی کند. در ابتدا، مبانی نظری و مفاهیم اصلی مرتبط با CD-CAT توضیح داده شده است. CD-CAT قابلیت ارائه ارزیابی دقیق‌تری از توانایی‌ها و مهارت‌های داوطلبان را دارد. این رویکرد با تحلیل نقاط قوت و ضعف در مهارت‌های خاص هر داوطلب، به‌جای ارائه یک نمره کلی، درک عمیق‌تری از وضعیت تحصیلی و توانمندی‌های فردی فراهم می‌کند. مقاله حاضر به تشریح مراحل مختلف ساخت آزمون با روش CD-CAT، از جمله طراحی بانک سؤال‌ها، الگوریتم‌های انتخاب سؤال و قواعد خاتمه آزمون می‌پردازد. همچنین، مزایای این رویکرد نسبت به آزمون‌های سنتی، از جمله افزایش دقت و پایایی ارزیابی‌ها و بهبود کارایی فرایندهای آموزشی، بررسی شده است. تحقیقات نشان می‌دهد که CD-CAT می‌تواند با استفاده از تحلیل پاسخ‌های قبلی داوطلبان، سؤال‌های متناسب با سطح توانایی آن‌ها را انتخاب کند و به‌این‌ترتیب، فرایند ارزیابی را بهینه سازد. همچنین یافته‌ها نشان می‌دهد که CD-CAT می‌تواند به بهبود کیفیت ارزیابی‌ها و افزایش دقت تشخیص‌ها کمک شایانی کند. در نهایت، پیشنهاداتی برای تحقیقات آینده در زمینه توسعه و بهبود CD-CAT و کاربردهای عملی آن در حوزه‌های مختلف آموزشی و ارزیابی ارائه شده است.

استناد: توانایی یوسفیان، س.، مینایی، ا.، عسگری، م.، شادرو، ش. و مشهدی، ع. (۱۴۰۴). سنجش انطباقی کامپیوتری تشخیصی شناختی (CD-CAT): چارچوب و گام‌های ساخت و رواسازی آزمون‌های آموزشی. *سنجش روانی تربیتی*، ۱(۱)، ۱۷۱-۱۹۱. https://doi.org/10.30470/ijpa.2025.726664

* **نویسنده مسئول:** سمیرا توانایی یوسفیان
نشانی: گروه سنجش و اندازه‌گیری، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبائی، تهران، ایران
پست الکترونیکی: tavanaei.yousefian@gmail.com

مقدمه

تحولات علمی و فناوری‌های نوین، اهمیت ارزیابی دقیق مشکلات یادگیری را در ارتقای کیفیت نظام آموزشی بیش‌ازپیش آشکار کرده‌اند. در این میان، روش‌های سنجش سنتی همچون آزمون‌های مداد – کاغذی، در پاسخ‌گویی به نیازهای متنوع آموزشی ناکارآمد ظاهر شده‌اند و این مسئله، لزوم بهره‌گیری از ابزارهای سنجش[1] پیشرفته را نمایان می‌سازد (آندراد مدیروس و مونیز[2]، ۲۰۲۲). سنجش آموزشی، تنها به اندازه‌گیری[3] دانش موجود محدود نمی‌شود، بلکه نقشی کلیدی در هدایت فرایند یادگیری، شناسایی نیازهای فردی و تقویت انگیزش ایفا می‌کند. ازاین‌رو، باید آن را به‌عنوان بخشی بنیادین از سیاست‌گذاری‌های آموزشی در نظر گرفت (عشیرت و اومور – سوبول[4]، ۲۰۲۴).

نظریه کلاسیک آزمون[5] (CTT) و نظریه سؤال – پاسخ[6] (IRT) به‌عنوان دو رویکرد رایج در سنجش صفات مکنون[7] آزمودنی‌ها، باوجود کاربرد گسترده، در تشخیص دقیق علل افت تحصیلی و نیازهای مهارتی فردی محدودیت‌هایی دارند. برای طراحی مداخلات آموزشی مؤثر، لازم است داده‌هایی دقیق‌تر درباره نیم‌رخ مهارتی[8] افراد گردآوری شود؛ امری که آزمون‌های سنتی در تحقق آن ناکارآمد هستند و این نارسایی می‌تواند اعتبار و دقت نتایج سنجش را تضعیف کند (سو[9]، ۲۰۲۲).

آزمون‌های سنتی خطی[10] که اغلب به‌صورت مداد – کاغذی اجرا می‌شوند، به‌علت یکسان‌بودن پرسش‌ها برای تمامی آزمودنی‌ها، در برآورد دقیق توانایی‌ها با محدودیت مواجه‌اند. افراد با توانایی پایین ممکن است از پاسخ‌گویی به سؤال‌های دشوار ناتوان باشند و افراد توانمند نیز با پرسش‌های ساده، سریع پیش روند. این پدیده منجربه کاهش دقت ارزیابی و نیاز به افزایش تعداد سؤال می‌شود (رودنر[11]، ۱۹۹۸). با پیشرفت فناوری، آزمون‌های مبتنی‌بر کامپیوتر[12] (CBT) توسعه یافتند که با ارائه بازخورد فوری، دقت بالا و کاهش خطاهای انسانی، جایگزینی مؤثر برای آزمون‌های سنتی محسوب می‌شوند (مگیس[13] و همکاران، ۲۰۱۷).

در قرن بیست و یکم، آزمون‌های مبتنی‌بر کاغذ[14] (PBT) با چالش‌های روزافزونی روبه‌رو شدند که ضرورت اصلاحات بنیادین را آشکار ساخت (بوساکوک[15] و همکاران، ۲۰۲۱). در پاسخ، آزمون‌های مبتنی‌بر کامپیوتر در قالب‌های متنوعی چون خطی، چندمرحله‌ای و انطباقی[16] توسعه یافتند و با بهره‌گیری از فناوری‌های چندرسانه‌ای، سطح دقت و کارایی بالاتری را ارائه دادند (مک کللند و کوواس[17]، ۲۰۲۰؛ جفری[18]، ۲۰۰۷؛ جئونگ[19]، ۲۰۱۴). پیشرفته‌ترین گونه این آزمون‌ها، سنجش انطباقی کامپیوتری[20] (CAT) است که با تنظیم سطح دشواری سؤال‌ها براساس توانایی آزمودنی، دقت و امنیت بیشتری را فراهم می‌آورد و به‌دلیل صرفه‌جویی در زمان و افزایش مشارکت، مورد توجه گسترده‌ای قرار گرفته است (بی[21] و همکاران، ۲۰۲۰؛ جفری، ۲۰۰۷). با بهره‌گیری از الگوریتم‌های هوشمند، سطح توانایی آزمودنی پس از هر پاسخ برآورد و سؤال بعدی براساس آن انتخاب می‌شود. نظریه سؤال پاسخ (IRT) و مدل‌های چندبعدی[22] (MIRT) به‌عنوان مبنای آماری این سیستم‌ها، امکان طراحی سؤال‌هایی با پارامترهای استاندارد نظیر دشواری[23] (b)، تمیز[24] (a)، حدس[25] (c) و برآورد لحظه‌ای توانایی (θ) را فراهم کرده‌اند (بی و همکاران، ۲۰۲۰).

فرایند CAT یک ساختار پویا، تعاملی و کاملاً فردمحور در سنجش تحصیلی است و مؤلفه‌هایی چون بانک سؤال‌های مدرج، الگوریتم انتخاب سؤال، برآورد پیوسته سطح توانایی، قواعد توقف آزمون و روش‌های نمره‌دهی نهایی را در بر می‌گیرد. فرایند CAT از پنج مؤلفه‌ی اصلی تشکیل شده است: ۱) قانون تصمیم‌گیری[26]، ۲) بانک سؤال[27]، ۳) شیوهٔ انتخاب سؤال[28]، ۴) شیوه نمره‌دهی یا برآورد توانایی[29]، ۵) قاعده اتمام آزمون[30] (تسائوسیس[31] و همکاران، ۲۰۲۱؛ بوساکوک و همکاران، ۲۰۲۱).

مطابق با شکل ۱، مراحل CAT به‌شرح زیر خلاصه می‌شوند:

گام اول (مرحله شروع): انتخاب سؤال‌های آغازین مناسب برای شروع آزمون؛

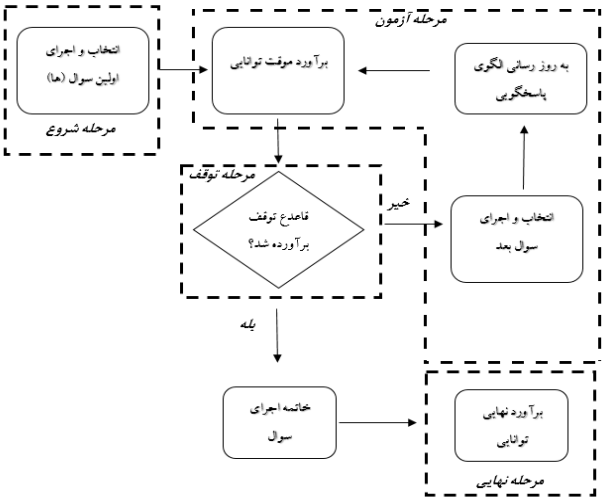
---

1 Assessment
2 Andrade Medeiros & Muniz
3 Measuring
4 Aşiret & Ömür-Sübül
5 Classical Test Theory (CTT)
6 Item Response Theory (IRT)
7 latent attributes
8 Students' skill profiles
9 Su
10 Linear Test
11 Rudner
12 Computer-Based Tests (CBT)
13 Magis
14 Paper Based Testing (PBT)
15 Boussakuk
16 adaptive
17 McClelland & Cuevas
18 Jafri
19 Jeong
20 Computerized Adaptive Testing (CAT)
21 Bi
22 Multidimensional item response theory (MIRT)
23 Difficulty
24 Discrimination
25 Guessing
26 Decision rule
27 Item pool
28 Item selection procedure / scoring procedure
29 Ability estimation method
30 Termination criterion
31 Tsaousis

گام دوم (مرحله اجرا): انتخاب پی‌درپی سؤال‌ها با به‌روزرسانی سطح توانایی پس از هر پاسخ؛

گام سوم (مرحله توقف): تعیین زمان اتمام آزمون بر اساس معیارهای توقف از پیش تعیین‌شده؛

گام چهارم (مرحله نهایی): برآورد نهایی توانایی و تولید گزارش‌های لازم (مگیس و همکاران، ۲۰۱۷).



شکل ۱
فرآیند CAT (مگیس و همکاران،۲۰۱۷)

CAT بر پایه نظریه IRT طراحی شده است که بر مفروضاتی نظیر تک‌بعدی بودن صفت مکنون و استقلال پاسخ‌ها متکی است (ویزمن[۱]، ۲۰۱۶). بااین‌حال، IRT بیشتر بر نمره کلی تمرکز دارد و قادر به تبیین دقیق تسلط آزمودنی‌ها بر مهارت‌های متنوع نیست (مینائی، ۱۳۹۱). در پاسخ به این کاستی، سنجش تشخیصی‌شناختی[۲](CDA) پدید آمده است که با ادغام روانسنجی و روان‌شناسی شناختی، تصویری دقیق‌تر از تسلط آزمودنی بر صفات خاص ارائه می‌دهد (جاویدان‌مهر و عنانی‌سراب، ۲۰۱۷).

در رویکرد CDA، تمرکز از نمره کل به‌سوی ترسیم نیم‌رخ‌های مهارتی تغییر یافته است. این نیم‌رخ‌ها ابزارهایی مؤثر برای طراحی آموزش‌های هدفمند محسوب می‌شوند (ویس و شاهین[۳]،۲۰۲۴). در مدل‌های تشخیصی‌شناختی[۴] (CDMs)، مهارت‌های مورد نیاز هر سؤال به‌صورت دوارزشی[۵] (۰ و ۱) در ماتریس کیو[۶]، ثبت می‌شوند. برای نمونه، اگر سؤالی نیازمند سه مهارت باشد، هشت ترکیب احتمالی ($2^3 = 8$) در پاسخ‌گویی محتمل است. ماتریس کیو که از تحلیل متخصصان حاصل می‌شود، ارتباط بین سؤال‌ها و مهارت‌های شناختی را مشخص می‌کند (وون داویر و لی[۷]، ۲۰۱۹).

مدل‌های CDM، برخلاف IRT که به‌صورت تک‌بعدی و پیوسته عمل می‌کند، از ساختاری چندبعدی و غیرخطی استفاده کرده و توانایی بیشتری در تحلیل تفاوت‌های فردی دارند (سسومز و هنسون[۸]، ۲۰۱۸؛ سو، ۲۰۲۲). از تلفیق قابلیت‌های CAT و CDM، رویکرد نوینی با عنوان سنجش انطباقی–کامپیوتری تشخیصی شناختی[۹] (CD-CAT) شکل گرفته است. این رویکرد، به‌جای صرف امتیازدهی، به شناسایی دقیق الگوهای چیرگی شناختی[۱۰] پرداخته و امکان طراحی آموزش‌های متناسب با نیازهای واقعی یادگیرنده را فراهم می‌کند (یانگ[۱۱]و همکاران،۲۰۲۰).

در سال‌های اخیر، CD-CAT به‌دلیل قابلیت ارائه بازخورد دقیق، شناسایی نقاط ضعف و ترسیم نیم‌رخ‌های مهارتی شخصی، مورد استقبال گسترده‌ای در حوزه آموزش قرار گرفته است. این رویکرد با ارتقای تصمیم‌گیری آموزشی مبتنی‌بر داده، به معلمان و طراحان در تدوین مداخلات مؤثر کمک می‌کند. برخلاف مدل‌های سنتی، CD-CAT نه‌تنها سطح تسلط را می‌سنجد، بلکه به‌صورت دقیق مشخص می‌کند آزمودنی در کدام مهارت‌ها نیاز به تقویت دارد و در کدام‌ها تسلط دارد (لین و چانگ[۱۲]، ۲۰۱۹).

1 Weissman
2 Cognitive Diagnostic Assessment (CDA)
3 Weiss & Sahin
4 Cognitive diagnosis models (CDMs)
5 Binary
6 Q-matrix
7 Von Davier & Lee
8 Sessoms & Henson
9 Cognitive diagnostic computerized adaptive testing (CD-CAT)
10 Cognitive Mastery Patterns
11 Yang
12 Lin & Chang

پژوهش حاضر با اتخاذ رویکردی مروری–تحلیلی، به تبیین ساختار و مؤلفه‌های کلیدی آزمون‌های مبتنی‌بر مدل‌های تشخیصی شناختی در بستر انطباقی کامپیوتری می‌پردازد. در این چارچوب، ساختار آزمون‌ها و همچنین منطق نظری حاکم بر طراحی آن‌ها مورد بررسی قرار می‌گیرد تا زمینه‌ای برای درک عمیق‌تر از سازوکارهای سنجش شناختی فراهم شود. هدف از این مطالعه، ارائه چارچوبی جامع برای طراحی آزمون‌هایی است که افزون بر سنجش سطح توانایی، به شناسایی نارسایی‌های یادگیری نیز بپردازند. با بهره‌گیری از پیشینه نظری معتبر در حوزه روانسنجی شناختی، این مقاله گامی مهم در جهت توسعه ابزارهای نوین ارزیابی آموزشی در کشور به‌شمار می‌رود و با تکیه بر مبانی نظری دقیق، زمینه‌ساز تصمیم‌گیری‌های آموزشی مبتنی‌بر شواهد خواهد بود.

سنجش انطباقی کامپیوتری تشخیصی شناختی (CD-CAT)

سنجش انطباقی کامپیوتری تشخیصی‌شناختی(CD-CAT) رویکردی نوین در حوزه روانسنجی است که تلفیقی از مدل‌های تشخیصی–شناختی (CDMs) و سنجش انطباقی کامپیوتری (CAT) را در یک چارچوب واحد ارائه می‌دهد. هدف این رویکرد، ارتقای دقت شناختی در ارزیابی‌های آموزشی و روانسنجی، هم‌زمان با بهره‌گیری از کارآمدی و انعطاف‌پذیری روش‌های انطباقی است (یو[1] و همکاران، ۲۰۱۹).

درCD-CAT، پاسخ‌های آزمودنی به‌صورت پی‌درپی تحلیل شده و نیم‌رخ چیرگی مهارت[2] او در هر مرحله به‌روزرسانی می‌شود. انتخاب سؤال‌های بعدی از بانک سؤال کالیبره‌شده به‌گونه‌ای انجام می‌شود که به بیشینه‌سازی آگاهی و کاهش تعداد سؤال‌های موردنیاز منجر گردد. این فرایند تا تحقق یکی از دو معیار توقف (تعداد سؤال‌ها یا سطح آستانه دقت[3] مطلوب) ادامه می‌یابد (لی[4] و همکاران، ۲۰۲۳).

برتری CD-CAT نسبت به CAT در پاسخ‌گویی به ساختارهای شناختی چندبعدی آزمودنی‌هاست. چنین ویژگی‌ای به‌ویژه در آزمون‌های تکوینی و حوزه‌هایی با ماهیت پیچیده و چندوجهی، مانند ریاضی و زبان، اهمیت بسزایی دارد (زو و چانگ[5]،۲۰۲۱).

در مدل‌های تشخیصی شناختی، ساختار سؤال‌ها از طریق ماتریس‌کیو مشخص می‌شود. این ماتریس، روابط بین سؤال‌های آزمون و صفات یا مهارت‌های شناختی را به‌صورت یک ماتریس K×D (K: تعداد سؤال‌ها، D: تعداد صفات) ترسیم می‌کند (راپ[6] و همکاران، ۲۰۱۰). اگر درایه q_{kd} برابر با ۱ باشد، به این معنا است که برای پاسخ‌دادن به سؤال k، تسلط بر صفت d ضروری است در غیر این صورت $q_{kd} = o$ که در آن: d =1, 2,.....,D و k =1, 2 ,.....,K

نیم‌رخ چیرگی مکنون برای آزمودنی n ام با α_n نشان داده می‌شود که مطابق بیان نمادین شماره (۱) به‌صورت زیر تعریف می‌گردد:

$$\alpha_n = (\alpha_{n1}, \alpha_{n2}, \dots\dots, \alpha_{nd}, \dots\dots, \alpha_{nD})' \qquad (۱)$$

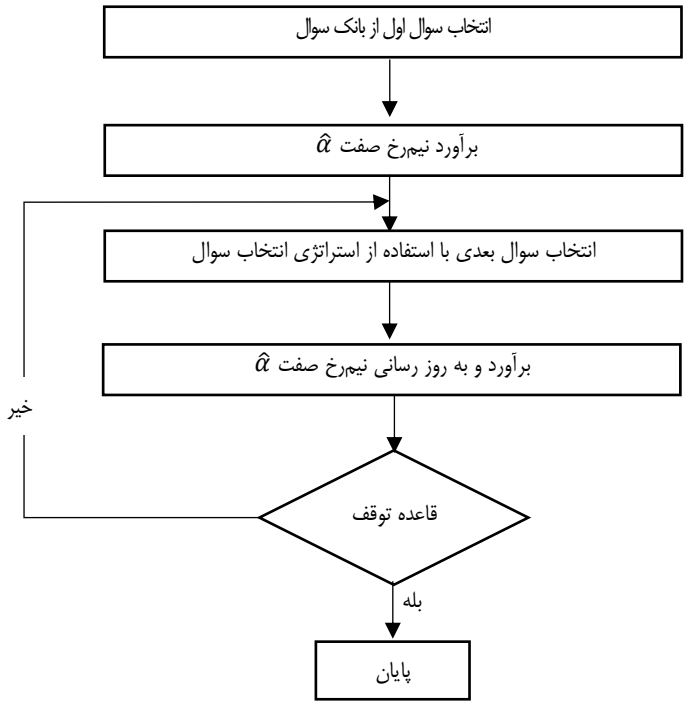
اگر n امین آزمودنی بر صفت d مسلط باشد $\alpha_{nd}=1$ در غیر این صورت، $\alpha_{nd}=0$ که n= 1,...,N است. N حجم کل نمونه است.

اصطلاح CD-CAT برای نخستین بار در مطالعه‌ای توسط چنگ[7] (۲۰۰۹) معرفی شد. این رویکرد، نوعی سنجش انطباقی کامپیوتری مبتنی‌بر مدل‌های تشخیصی شناختی (CDMs) است که با هدف ارزیابی دقیق‌تر صفات شناختی آزمودنی‌ها طراحی شده و به‌دلیل توان بالای تطبیق با صفت‌های فردی و دقت بالا در تشخیص، به‌سرعت مورد توجه روانسنجان قرار گرفت. روند کلی اجرای CD-CAT در شکل ۲ به‌صورت فلوچارت نمایش داده شده است و مراحل آزمون را از انتخاب سؤال اول تا توقف بر اساس معیارهایی چون دقت[8] یا تعداد سؤال نشان می‌دهد. پس از انتخاب نخستین سؤال، پاسخ آزمودنی به‌منظور برآورد اولیه نیم‌رخ چیرگی شناختی تحلیل می‌شود و در ادامه، با هر پاسخ جدید، این نیم‌رخ به‌روزرسانی شده و سؤال بعدی انتخاب می‌شود. این فرایند تا تحقق معیار توقف ادامه می‌یابد و در نهایت، برآوردی دقیق از صفت‌های شناختی فرد ارائه می‌گردد.

در رویکرد سنجش انطباقی‌کامپیوتری تشخیصی شناختی (CD-CAT)، هدف اصلی، برآورد نیم‌رخ چیرگی مکنون[9] آزمودنی است؛ یعنی تشخیص اینکه فرد در کدام صفات یا مهارت‌های شناختی تعریف‌شده، تسلط دارد یا ندارد. در نسخه‌های پیشرفته‌تر CD-CAT، افزون بر نیم‌رخ چیرگی شناختی (α)، توانایی کلی آزمودنی (θ) نیز به‌طور هم‌زمان تخمین زده می‌شود. در این صورت، اگر تنها α ارزیابی شود، آزمون تک‌منظوره[10] و در صورت سنجش هم‌زمان α وθ ، آزمون دومنظوره[11] نامیده می‌شود (وون و همکاران، ۲۰۱۹).

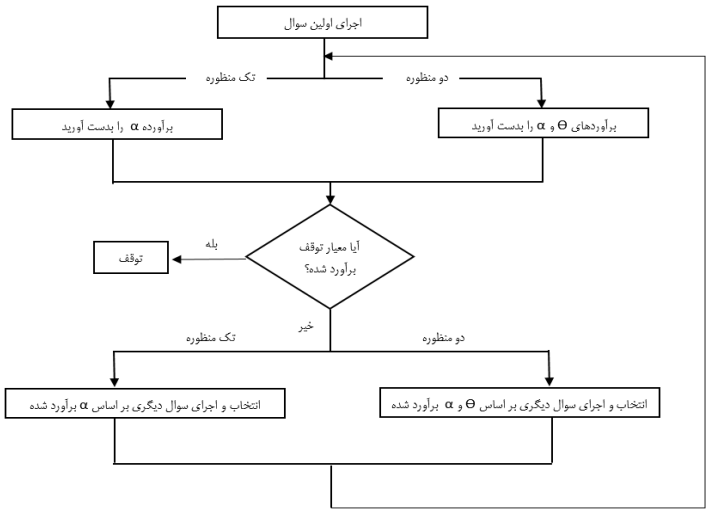
فرآیند اجرای CD-CAT در قالب یک فلوچارت (شکل ۳) نمایش داده شده است.

[1] Yu
[2] Attribute mastery profile
[3] Precision cut
[4] Li
[5] Zhu & Chang
[6] Rupp
[7] Cheng
[8] Precision
[9] Latent mastery profile
[10] Single purpose
[11] Dual purposes



شکل۲: فرآیند کلی اجرای CD-CAT (لی و همکاران، ۲۰۲۳)

همان‌طور که از تصویر ۳ مشخص است در فرآیند اجرای روش CD-CAT، انتخاب سؤال براساس نیم‌رخ چیرگی (α) و در حالت دومنظوره، براساس θ نیز قابل انجام است. پس از پاسخ به اولین سؤال، α (و در صورت لزوم θ) برآورد می‌شود و با تحلیل هر پاسخ، مقادیر به‌روزرسانی شده و سؤال بعدی انتخاب می‌گردد. این انتخاب می‌تواند با هدف افزایش دقت تشخیص صفات شناختی یا براساس الگوهای پاسخ مشابه آزمودنی‌ها صورت گیرد. این فرایند تا تحقق یکی از معیارهای توقف، مانند رسیدن به دقت مطلوب یا تعداد مشخصی سؤال، ادامه پیدا می‌کند. CD-CAT با تطبیق آنی و فردمحور سؤال‌ها با ویژگی‌های شناختی آزمودنی، تجربه‌ای شخصی‌سازی‌شده، دقیق و انسان‌محور فراهم می‌سازد؛ در عین‌حال که کارایی روانسنجی نیز حفظ می‌شود.



شکل ۳: الگوریتم انتخاب سؤال در CD-CAT بر پایه α و θ (وون و همکاران، ۲۰۱۹)

فرآیند سنجش انطباقی کامپیوتری تشخیصی شناختی (CD-CAT)

CAT روشی فردمحور است که با هدف شناسایی صفات شناختی پنهان آزمودنی طراحی شده و با انتخاب نخستین سؤال از بانک پرسش‌ها آغاز می‌شود. پس از تحلیل پاسخ، نیمرخ مهارتی[1] اولیه (α) برآورد می‌شود و سؤال بعدی براساس این تخمین انتخاب می‌گردد. این چرخه تا زمان تحقق معیارهای توقف مانند سطح اطمینان کافی یا حداکثر تعداد سؤال‌ها ادامه می‌یابد. در الگوریتم‌های ساده، انتخاب سؤال تنها بر پایه α (مدل تک‌منظوره)

[1] Skill Profile

انجام می‌شود، اما در نسخه‌های پیشرفته‌تر، همزمان از α و θ استفاده می‌شود (مدل دومنظوره) تا دقت و کارایی افزایش یابد. این سازوکار، همسو با هدف اصلی CD-CAT یعنی درک عمیق و دقیق ساختارهای شناختی آزمودنی است (وون، داویر و لی، ۲۰۱۹).

CD-CAT بر مبنای ترکیب نظریه سؤال-پاسخ (IRT) و مدل‌های تشخیصی شناختی (CDMs) طراحی شده و با چارچوب‌های رایج در سنجش انطباقی (CAT) هم‌راستاست؛ به‌ویژه در هدف‌گیری برای طبقه‌بندی دقیق و کاهش طول آزمون با بیشینه‌سازی آگاهی هر سؤال. مطابق با چارچوب پنج‌مرحله‌ای مطرح‌شده توسط هوبنر[1] (۲۰۱۰) و توسعه‌یافته در بررسی فنی یو و همکاران (۲۰۱۹)، اجرای CD-CAT شامل پنج گام اصلی است: ۱- بانک سؤال‌ها، ۲- قاعده شروع[2]، ۳- روش انتخاب سؤال[3] (ISS)، ۴- روش نمره‌دهی[4]، ۵- قاعده توقف[5] (ویس و شاهین، ۲۰۲۴).

۱- بانک سؤال

بانک سؤال در سنجش انطباقی‌کامپیوتری‌تشخیصی شناختی(CD-CAT) همانند CAT، نقش بنیادی در دقت و اثربخشی فرآیند سنجش ایفا می‌کند؛ با این تفاوت اساسی که در CD-CAT، این بانک به‌شدت به ماتریس کیو وابسته است. در واقع، CD-CAT برای استنتاج نیمرخ شناختی آزمودنی‌ها، نه‌تنها به پاسخ‌های آن‌ها بلکه به رابطه دقیق میان سؤال‌ها و مهارت‌های شناختی زیربنایی متکی است؛ رابطه‌ای که به‌طور صریح در ماتریس کیو بازنمایی می‌شود.کیفیت و دقت ماتریس کیو نقشی اساسی و حیاتی در صحت و روایی این فرایند ایفا می‌کند. ماتریس کیو به‌صورت یک ماتریس دودویی طراحی می‌شود که مشخص می‌کند پاسخ صحیح به هر سؤال، مستلزم تسلط بر کدام مهارت‌ها یا صفات شناختی است. این ماتریس معمولاً بر اساس قضاوت جمعی متخصصان موضوعی[6] و با بهره‌گیری از تحلیل‌های محتوایی تهیه می‌شود. دقت و اعتبار این ماتریس از اهمیت ویژه‌ای برخوردار است؛ زیرا هرگونه خطا یا کاستی در نگاشت مهارت‌ها می‌تواند به استنتاج‌های نادرست درباره نیمرخ شناختی آزمودنی منجر شود و در نهایت روایی آزمون را تهدید کند (راپ و تمپلین[7]، ۲۰۰۸؛ یو و همکاران، ۲۰۱۹).

CD-CAT با بهره‌گیری از مدل‌های تشخیصی شناختی (CDMs) نظیر DINA[8]، DINO[9] و G-DINA[10]، علاوه‌بر اطلاعات آماری، ساختار شناختی آزمودنی را نیز در فرآیند انتخاب سؤال لحاظ می‌کند. برای مثال، مدل DINA یک مدل محدودکننده محسوب می‌شود که فرض می‌کند فرد تنها زمانی می‌تواند به یک سؤال پاسخ صحیح دهد که بر همه مهارت‌های لازم تسلط کامل داشته است (ویس و شاهین،۲۰۲۴).

از آن‌جا که CD-CAT رویکردی نوین در حوزه روانسنجی به شمار می‌آید، منابع علمی اندکی به‌طور مستقیم به ویژگی‌های مطلوب بانک سؤال در این نوع آزمون پرداخته‌اند. چنگ (۲۰۰۹) اشاره می‌کند که به‌دلیل پیچیدگی طراحی ماتریس کیو، احتمال آن‌که حجم بانک سؤال در CD-CAT با بانک‌های مرسوم در CAT برابر باشد، پایین است؛ چرا که طراحی هر سؤال مستلزم تعریف دقیق نگاشت مهارت‌ها بوده و این فرایند نیازمند صرف زمان و تخصص بالاست. بنابراین، انتظار برخورداری CD-CAT از بانک سؤال‌های با حجمی مشابه آزمون‌های CAT ، چندان واقع‌بینانه نیست. در حالی که CAT ممکن است شامل ۱۰۰ سؤال یا بیشتر باشد، بانک سؤال در CD-CAT معمولاً محدودتر است (برای مثال، بین ۳۰ تا ۵۰ سؤال)، بدون آن که لزوماً منجر به کاهش دقت طبقه‌بندی شناختی شود. این ویژگی ناشی از انتخاب هدفمند و مبتنی بر مهارت سؤال‌ها در فرآیند تطبیق نهفته است. به عبارت دیگر، CD-CAT ضمن بهره‌گیری از اطلاعات آماری، ساختار شناختی آزمودنی را نیز در انتخاب سؤال بعدی لحاظ می‌کند؛ همین امر به افزایش کارآمدی فرآیند ارزیابی منجر می‌شود.

طراحی ماتریس کیو فرآیندی نظام‌مند و چندمرحله‌ای علمی است که باید با مشارکت فعال متخصصان حوزه محتوا، روانسنجان و طراحان سؤال انجام شود. این فرآیند در چهار گام اصلی در ادامه توضیح داده می‌شود.

گام اول؛ شناسایی و تعریف مهارت‌های شناخت: در این مرحله، مجموعه‌ای از مهارت‌های مرتبط با محتوای آزمون استخراج و به‌صورت دقیق، رفتاری و قابل مشاهده تعریف می‌شود. این مهارت‌ها باید بر مبنای تحلیل محتوای درسی، اهداف آموزشی، و نظر کارشناسان آموزشی تعیین شوند (راپ و همکاران،۲۰۱۰). به‌عنوان نمونه، در سطح دبستان و در آزمونی برای سنجش مهارت‌های پایه ریاضی، می‌توان مهارت‌ها را به این‌صورت تعریف کرد (ویس و شاهین، ۲۰۲۴). مهارت ۱: جمع؛ مهارت ۲: تفریق؛ و مهارت ۳: ضرب.

گام دوم؛ نگاشت سؤال‌ها به مهارت‌ها (ساخت ماتریس کیو): هر سؤال باید از نظر محتوایی تحلیل شود تا مشخص گردد پاسخ صحیح به آن، مستلزم بهره‌گیری از کدام‌یک از مهارت‌های تعریف‌شده در چارچوب نظری آزمون است. انجام این فرایند مستلزم همکاری مشترک میان متخصصان موضوعی(SMEs) و روانسنجان است تا اعتبار محتوایی ماتریس کیو به‌درستی تأمین شود (راپ و تمپلین، ۲۰۰۸). به‌عنوان نمونه، جدول ۱ نمایی از یک ماتریس کیو را برای آزمونی شامل ۱۰ سؤال و ۳ مهارت شناختی نشان می‌دهد (ویس و شاهین، ۲۰۲۴).

---

[1] Huebner
[2] Starting Rules
[3] Item Selection Strategies
[4] Scoring rule
[5] Stopping Rules
[6] Subject Matter Experts
[7] Rupp and Templin
[8] Deterministic Inputs, Noisy “And” gate model (DINA)
[9] Deterministic Inputs, Noisy “Or” gate model (DINO)
[10] Generalized Deterministic Inputs, Noisy “And” gate model (G-DINA)

جدول ۱
ماتریس‌کیو برای یک آزمون ریاضی شامل ۱۰ سؤال و ۳ مهارت شناختی

سؤال	مهارت ۱ (جمع)	مهارت ۲ (تفریق)	مهارت ۳ (ضرب)
۱	۱	۰	۱
۲	۱	۰	۱
۳	۱	۱	۱
۴	۱	۰	۰
۵	۱	۱	۰
۶	۱	۱	۱
۷	۱	۰	۱
۸	۱	۱	۱
۹	۱	۱	۰
۱۰	۱	۱	۰

برای روشن‌تر شدن موضوع، نمونه‌ای از سؤال‌ها به همراه نگاشت مهارتی آن‌ها به‌شرح زیر ارائه می‌گردد:

سؤال ۱: (۲ + ۴) + (۶ × ۸) = ؟ ⟵ نیازمند مهارت ۱ (جمع) و ۳ (ضرب)

سؤال ۵: (۳۱ + ۱۶) + (۱۷ − ۱۱) = ؟ ⟵ نیازمند مهارت ۱ (جمع) و ۲ (تفریق)

سؤال ۱۰: (۱۲ − ۸) + (۱۳ × ۴) = ؟ ⟵ نیازمند مهارت‌های ۱، ۲ و ۳

همان‌طور که در مثال‌ها مشاهده می‌شود، مهارت ۱ (جمع) در تمامی سؤال‌های آزمون مشترک است و به‌عنوان یک مهارت محوری در ساختار کلی آزمون ایفای نقش می‌کند. این ویژگی نشان می‌دهد که طراحی آزمون باید به‌گونه‌ای صورت گیرد که نه تنها پوشش کامل و هدفمندی از این مهارت فراهم شود، بلکه سایر مهارت‌ها نیز در قالب ترکیب‌های متنوع و متوازن در سؤال‌ها لحاظ شوند.

در واقع، تکرار حضور مهارت ۱ در کلیه سؤال‌ها، زنگ هشداری برای طراحان آزمون است تا نسبت به توزیع یکنواخت مهارت‌ها حساسیت بیشتری نشان دهند. توجه به تعادل شناختی میان مهارت‌ها، نه‌تنها از منظر روانسنجی ضروری است، بلکه نقش تعیین‌کننده‌ای در تفسیرهای تشخیصی مبتنی بر عملکرد آزمودنی‌ها دارد.

با توجه به آن‌که داده‌های ماتریس‌کیو، پایه‌ی اصلی مدل‌های تشخیصی شناختی (CDMs) را تشکیل می‌دهد، تعیین دقیق روابط میان سؤال‌ها و مهارت‌ها اهمیت دوچندان می‌یابد. از این‌رو، وجود اجماع میان متخصصان موضوعی (SMEs) درباره مهارت‌های لازم برای پاسخ‌دهی به هر سؤال، شرط اساسی برای اعتبار نتایج به‌دست‌آمده از این مدل‌ها به‌شمار می‌رود.

گام سوم؛ اعتبارسنجی و تحلیل روایی ماتریس کیو: برای اطمینان از دقت نگاشت سؤال‌ها به مهارت‌های شناختی، ضروری است که ماتریس‌کیو توسط چند ارزیاب مستقل مورد بررسی قرار گیرد. میزان توافق میان ارزیاب‌ها را می‌توان با بهره‌گیری از شاخص‌های آماری نظیر ضریب کاپای کوهن[1] اندازه‌گیری کرد که نشان‌دهنده‌ی میزان هم‌پوشانی قضاوت‌های آن‌ها فراتر از شانس است (کوهن،۱۹۶۰). در ادامه و پس از اجرای نسخه‌ی آزمایشی آزمون، داده‌های به‌دست‌آمده با استفاده از مدل‌های تشخیصی شناختی (CDMs) نظیر مدل DINA یا G-DINA مورد تحلیل قرار می‌گیرند. تحلیل‌های انجام‌شده در این مرحله، به شناسایی نواقص احتمالی در ساختار ماتریس کیو کمک می‌کند و زمینه را برای اعمال اصلاحات لازم فراهم می‌سازد. این بازبینی و اصلاح، گامی حیاتی در جهت افزایش روایی تشخیصی ابزار اندازه‌گیری محسوب می‌شود (دلاتوره[2]، ۲۰۰۹).

گام چهارم؛ طبقه‌بندی آزمودنی‌ها بر اساس پاسخ‌ها: در مرحله نهایی، براساس پاسخ‌های ثبت‌شده و تحلیل آن‌ها توسط مدل تشخیصی، هر آزمودنی در یکی از طبقات مکنون قرار می‌گیرد (راپ و همکاران، ۲۰۱۰). برای سه مهارت، هشت ترکیب ممکن ($2^3 = 8$) از حالت‌های تسلط یا عدم تسلط (از ۰۰۰ تا ۱۱۱) وجود دارد که بر اساس آن، به هر آزمودنی یک برچسب طبقه‌ای اختصاص داده می‌شود. در جدول ۲، نمونه‌ای فرضی از احتمال‌های برآوردشده برای یک آزمودنی ارائه شده است، که احتمال تعلق وی به هر یک از این طبقات را نشان می‌دهد. همچنین برای وضوح بیشتر، ردیف شماره‌گذاری طبقات با عنوان مشخص شده است.

[1] Cohen's Kappa coefficient

[2] de la Torre

جدول ۲

الگوهای تسلط و توزیع احتمالات طبقه‌بندی آزمودنی‌ها در ماتریس کیو

الگو	شماره طبقه							
	۱	۲	۳	۴	۵	۶	۷	۸
الگوی تسلط	۰۰۰	۰۰۱	۰۱۰	۱۰۰	۰۱۱	۱۰۱	۱۱۰	۱۱۱
احتمالات آزمودنی	۰/۱۲	۰/۰۶	۰/۰۸	۰/۰۵	۰/۴۸	۰/۱۲	۰/۰۷	۰/۰۲

در این جدول، عدد «۱» نشان‌دهنده تسلط و عدد «۰» نمایانگر عدم تسلط بر هر مهارت است. بر این اساس، این آزمودنی بیشترین احتمال را برای قرارگیری در طبقه ۴ دارد؛ یعنی عدم تسلط بر مهارت‌های ۱ و ۲ و تسلط بر مهارت ۳. با توجه به داده‌های جدول، احتمال تسلط دانش‌آموز بر مهارت ۱ برابر با مجموع احتمالات تمامی طبقه‌هایی است که در الگوی آن‌ها، مهارت یک با «۱» مشخص شده است. در این مثال، طبقه‌های ۲، ۵، ۶ و ۸ شامل مهارت ۱ هستند؛ بنابراین، احتمال تسلط دانش‌آموز بر این مهارت برابر با ۰/۶۸ است:

۰/۰۶ + ۰/۴۸ + ۰/۱۲ + ۰/۰۲ =۰/۶۸

همچنین، احتمال تسلط بر مهارت ۲ برابر با ۰/۶۵ محاسبه می‌شود که مجموع احتمالات طبقه‌های ۳، ۵، ۷ و ۸ است:

۰/۰۸ + ۰/۴۸ + ۰/۰۷ + ۰/۰۲ = ۰/۶۵

و تسلط بر مهارت ۳ برابر با ۰/۲۶ است.

۰/۰۵ + ۰/۱۲ + ۰/۰۷ + ۰/۰۲ = ۰/۲۶

طبقه‌بندی آزمودنی‌ها بر پایه مدل‌های تشخیصی شناختی، گامی بنیادین در شناسایی دقیق شایستگی‌ها و تحلیل نظام‌مند نقاط قوت و ضعف شناختی هر فرد به شمار می‌آید. در این راستا، طراحی علمی و اعتبارسنجی دقیق بانک سؤال و ماتریس کیو، به‌ویژه در چارچوب آزمون‌های تطبیقی شناخت‌محور (CD-CAT)، نه‌تنها دقت شناسایی الگوهای تسلط مهارتی را افزایش می‌دهد و زمینه‌ساز درکی ژرف‌تر از فرآیندهای ذهنی و یادگیری آزمودنی می‌شود، بلکه به‌واسطه ماهیت تطبیقی آزمون، امکان ارائه مسیرهای منحصربه‌فرد و شخصی‌سازی‌شده برای هر فرد را نیز فراهم می‌سازد (ویس و شاهین، ۲۰۲۴).

۲- قاعده شروع

یکی از روش‌های رایج برای آغاز فرایند آزمون CD-CAT، استفاده از مجموعه‌ای مشترک از سؤال برای تمامی آزمون‌شوندگان است. سپس بر اساس پاسخ به این سؤال‌ها، احتمال عضویت در هر طبقه چیرگی[1] برای هر دانش‌آموز برآورد شده و از این احتمال‌ها برای آغاز CD-CAT به‌صورت سؤال‌به‌سؤال استفاده می‌شود (ویس و شاهین،۲۰۲۴). زو[2] و همکاران (۲۰۰۳) در یک مطالعه شبیه‌سازی، پنج سؤال مشابه را به‌طور تصادفی به هر شرکت‌کننده ارائه کردند. در حالی که لیو[3] و همکاران (۲۰۱۳) در یک مطالعه واقعی CD-CAT از هشت سؤال اولیه برای برآورد احتمال‌های شروع آزمون استفاده کردند. روش‌های دیگر پیشنهادی برای شروع CD-CAT شامل انتخاب تصادفی سؤال‌ها، اختصاص یک نیم‌رخ چیرگی به‌صورت تصادفی (چن[4] و همکاران، ۲۰۱۲)، یا استفاده از اطلاعات دموگرافیک حاصل از پیش‌آزمون‌ها برای انتخاب سؤال‌آغازین است (وون داویر و چنگ[5]، ۲۰۱۴). همچنین اگر عملکرد فرد به‌صورت طولی دنبال شود، نیم‌رخ چیرگی قبلی می‌تواند به‌عنوان نقطه شروع آزمون بعدی به‌کار رود (وانگ[6]، ۲۰۲۱).

به لحاظ نظری، یک سیستم CAT می‌تواند آزمون را در هر سطحی از دشواری آغاز کند. ساده‌ترین قاعده آن است که آزمون با یک سؤال یکسان یا با فرض سطح توانایی اولیه برابر با صفر برای همه آزمون‌دهندگان شروع شود. سپس، انتخاب سؤال بعدی بر اساس الگوریتم انتخاب سؤال انجام می‌گیرد. در CD-CAT نیز می‌توان شروع آزمون را بر مبنای رایج‌ترین نیم‌رخ چیرگی در جامعه در نظر گرفت.

با این حال، استفاده از سؤال‌های ثابت و پرتکرار در ابتدای آزمون ممکن است به آشنایی آزمون‌دهندگان با سؤال‌ها منجر شود و اعتبار آزمون را تحت تأثیر قرار دهد. به‌همین دلیل، معمولاً عنصر تصادفی‌سازی در قاعده شروع گنجانده می‌شود. به‌عنوان مثال، یکی از روش‌های متداول، شروع آزمون با انتخاب تصادفی یک مقدار اولیه از θ در بازه‌ای خاص، (مثلا (۰/۵، ۰/۵-) است؛ یا شروع آزمون با انتخاب تصادفی یک سؤال یا چند سؤال از میان مجموعه‌ای از سؤال‌های از پیش تعیین‌شده با سطح دشواری متوسط است (وون داویر و لی، ۲۰۱۹).

[1] mastery class
[2] Xu
[3] Liu
[4] Chen
[5] Davier & Cheng
[6] Wang

رایلی[1] و همکاران (۲۰۰۷) نیز اجرای یک پیش‌آزمون کوتاه را برای کسب اطلاعات اولیه جهت شروع آزمون پیشنهاد کردند. در CD-CAT، برخی محققان بر شروع آزمون با سؤال‌های تصادفی و یکسان برای همه آزمون‌دهندگان تأکید دارند (زو و همکاران، ۲۰۰۳)؛ در حالی که برخی دیگر اختصاص تصادفی یک نیم‌رخ چیرگی و سپس انتخاب سؤال‌های هماهنگ با آن با استفاده از روش‌هایی نظیر روش انتخاب سؤال (ISS)، را مناسب می‌دانند (چن و همکاران، ۲۰۱۲). در نتیجه، می‌توان گفت که انتخاب سؤال‌های آغازین در CD-CAT می‌تواند از میان مجموعه‌ای از سؤال‌های استاندارد، یا به‌صورت تصادفی از کل بانک سؤال انجام شود (وون داویر و لی، ۲۰۱۹).

۳- روش انتخاب سؤال

در CD-CAT، اجزایی مانند استراتژی انتخاب سوا ل[2](ISS) بسته به هدف می‌توانند ساختار متفاوتی نسبت به CAT سنتی داشته باشند، زیرا در CD-CAT ارزیابی بر اساس طبقات مکنون گسسته انجام می‌شود، نه صفات پیوسته. به‌همین دلیل، بسیاری از الگوریتم‌های ISS سنتی، مانند اطلاعات فیشر حداکثری[3] (MFI)، برای CD-CAT مناسب نیستند و عملکرد دقیقی در شرایط با تعداد سؤال کم ندارند (عشیرت و اومور–سوبول، ۲۰۲۴). بنابراین، الگوریتم‌های جدیدی مخصوص CD-CAT توسعه یافته‌اند. هدف ISS درCD-CAT، برآورد دقیق نیم‌رخ چیرگی (α) یا ترکیب α و θ است، در حالی که ملاحظاتی مانند کنترل محتوایی و میزان مواجهه[4] با سؤال نیز مدنظر قرار می‌گیرند. این الگوریتم‌ها به دو دسته پارامتری و ناپارامتری تقسیم می‌شوند (وون، داویر و لی، ۲۰۱۹).

تمرکز اصلی پژوهش‌ها در CD-CAT، مشابه CAT، بر بهینه‌سازی انتخاب سؤال است؛ زیرا این فرآیند نقش حیاتی در افزایش کارایی آزمون دارد. درCAT، پرسش‌ها با بیشترین میزان آگاهی انتخاب می‌شوند تا دقت بالا و خطای معیار اندازه گیری[5] (SEM) پایین حاصل شود. اگرچه در CD-CAT معادل مستقیمی برای SEM وجود ندارد، اما هدف نهایی همچنان دستیابی به تخصیص دقیق و معتبر آزمون‌شوندگان به طبقات مکنون بر اساس الگوی پاسخ آن‌هاست. در این راستا، دو روش عمده انتخاب سؤال درCD-CAT، روش کولبک–لایبلر[6] (KLI) و آنتروپی شانون[7] (SHE) هستند (ویس و شاهین، ۲۰۲۴؛ هوبنر، ۲۰۱۰).

روش کولبک–لایبلر (KLI) فاصله بین دو توزیع احتمال را می‌سنجد و درCD-CAT، تفاوت بین احتمال‌های طبقه‌بندی فعلی و سایر طبقات مکنون را برای هر سؤال باقی‌مانده در بانک سؤال اندازه‌گیری می‌کند. سؤال با بیشترین مقدار KLI انتخاب می‌شود؛ زیرا بیشترین تمایز را در طبقه‌بندی ایجاد می‌کند. از سوی دیگر، SHE نشانگر میزان عدم قطعیت در توزیع احتمال طبقات مکنون است. درCD-CAT، سؤالی انتخاب می‌شود که کمترین مقدار SHE را داشته باشد، چرا که این سؤال بیشترین اطمینان را در پیش‌بینی نیم‌رخ چیرگی آزمودنی فراهم می‌سازد. زمانی که یکی از طبقات مکنون احتمال ۱.۰ و سایر طبقات احتمال ۰.۰ داشته باشند، مقدار SHE حداقل است. در مقابل، اگر احتمال‌ها یکسان باشند، SHE حداکثر می‌شود.

چنگ (۲۰۰۹) عملکرد نسخه‌های پایه و بهبودیافته KLI و SHE را مقایسه کرد و نشان داد که نسخه‌های بهبودیافته دقت بالاتری دارند. یو و همکاران (۲۰۱۹) حوزه انتخاب سؤال را گسترش دادند و چندین روش جدید برای مراحل اولیهCD-CAT ارائه کردند، با هدف کاهش خطای طبقه‌بندی زمان‌که تعداد سؤال‌ها اندک است. آن‌ها همچنین انتخاب سؤال‌هایی را بررسی کردند که امکان برآورد همزمان احتمال عضویت در طبقات مکنون و برآورد θ را فراهم می‌کنند؛ به‌عبارت‌دیگر، اجرای همزمان CD-CAT و CAT را ممکن می‌سازند. مانند CAT، در این پژوهش‌ها ملاحظاتی همچون تعادل محتوایی نیز در نظر گرفته شده است (ویس و شاهین، ۲۰۲۴).

ISS های طراحی‌شده برای CD-CAT را می‌توان به سه دسته کلی تقسیم کرد: تک منظوره برای برآورد α، دومنظوره برای برآورد α وθ، و برخی از روش‌های پیشرفته انتخاب سؤال در کنار معیارهای آماری، محدودیت‌های مفهومی مانند تنوع در سطوح شناختی یا پوشش متوازن مفاهیم را نیز درنظر می‌گیرند. از جمله شاخص‌های شناخته‌شده می‌توان به شاخص‌های کولبک – لایبلر با وزن پیشین اصلاح شده[8] (MPWKL) و شاخص کولبک – لایبلر با وزن پسین[9] (PWKL) و شاخص تمایز مدل تعمیم‌یافته[10] (GDI) اشاره کرد (وون داویر و لی، ۲۰۱۹).

به باور لو و همکاران[11] (۲۰۲۳)، پژوهش‌ها در زمینه انتخاب سؤال عمدتاً دو جنبه دارند: (۱) دقت طبقه‌بندی، شامل روش‌هایی چون: روش آنتروپی شانون[12] (SHE) (زو و همکاران، ۲۰۰۳)، روش کولبک–لایبلر با وزن پسین[13] (PWKL) (چنگ، ۲۰۰۹)، روش اطلاعات متقابل[14] (MI) (وانگ[15]،

[1] Riley
[2] Item Selection Strategies ISS
[3] Maximum Fisher Information (MFI)
[4] exposure
[5] standard error of measurement (SEM)
[6] Kullback-Leibler
[7] Shannon entropy
[8] Modified posterior-weighted Kullback–Leibler index (MPWKL)
[9] Posterior weighted Kullback–Leibler (PWKL)
[10] Generalized Model Discrimination Index (GDI) Method
[11] Luo et al
[12] Shannon Entropy (SHE) Method
[13] Posterior-weighted Kullback–Leibler information (PWKL) method
[14] Mutual information (MI) method
[15] Wang

۲۰۱۳)، روش اصلاح شده کولبک–لایبلر با وزن[1] (MPWKL)، روش شاخص تمیز مدل تعمیم یافته[2] (GDI) (کاپلان [3]و همکاران،۲۰۱۵)، روش‌های مبتنی بر شاخص تمییز[4] (CDI) (ژنگ و چانگ[5]، ۲۰۱۶)؛ و همچنین روش‌هایی برای متوازن‌سازی صفت‌ها[6] (وانگ[7] و همکاران، ۲۰۲۰)؛ (۲) حفظ دقت طبقه‌بندی در طول آزمون همراه با بهینه‌سازی نرخ استفاده از سؤال‌ها، شامل: طراحی متعادل کننده صفت[8] (لین و چانگ، ۲۰۱۹)، طراحی طبقه‌بندی شده[9] (یانگ و همکاران، ۲۰۲۰)، طراحی مسدود شده[10] (کاپلان و دلا تورە[11]، ۲۰۲۰) و جستجوی دودویی[12] (ژنگ و وانگ[13]، ۲۰۱۷).

کاپلان و همکاران (۲۰۱۵) شاخص GDI را پیشنهاد کردند که واریانس وزنی احتمال پاسخ صحیح به یک سؤال را با توجه به توزیع طبقات مهارتی آزمودنی‌ها مدنظر قرار می‌دهد. با توجه به اینکه شاخص GDI کمترین زمان محاسبه را در مقایسه با دو شاخص دیگر (JSD[14] و PWKL) دارد(کاپلان و همکاران، ۲۰۱۵)، برای کاربرد در آزمون‌های انطباقی کوتاه بسیار مناسب است.

با توجه به اینکه شاخص تمایز مدل تعمیم‌یافته (GDI) در مقایسه با شاخص واگرایی جنسن–شانون (JSD) و روش کولبک–لایبلر با وزن پسین (PWKL)، زمان محاسبه‌ی بسیار کمتری دارد (کاپلان و همکاران، ۲۰۱۵)، می‌توان آن‌را گزینه‌ای مناسب برای استفاده در آزمون‌های انطباقی با طول کوتاه دانست. به‌طور مشخص، در شاخص GDI برای سؤال j در چرخه (مرحله) t آزمون، مقدار واریانس ς_j^2 مطابق فرمول شماره (۲) فرمول‌بندی می‌شود.

$$\varsigma_j^2 = \sum_{=1}^{2^{k_j^*}} (P(Y_j = 1|\boldsymbol{\alpha}_{lj}^*) - \bar{P}_j)^2 \pi^{(t)}(\boldsymbol{\alpha}_{lj}^*|\boldsymbol{y}^{(t)}) \quad (۲)$$

که در آن:

j اندیس سؤال است، یعنی سؤال jام که در آزمون مطرح می‌شود؛

T شماره چرخه یا مرحله آزمون است؛ هر چرخه برابر با ارائه یک سؤال و دریافت پاسخ است؛

k_j^* عداد مهارت‌های مرتبط با سؤال j است (تعداد ابعاد صفات برای آن سؤال)؛

l اندیس طبقه مهارتی است که از ۱ تا $2^{k_j^*}$ متغیر است، زیرا تعداد کل طبقات مهارتی برابر با توان ۲ تعداد مهارت‌های مرتبط با سؤال است؛

$(P(Y_j = 1|\boldsymbol{\alpha}_{lj}^*)$ احتمال پاسخ صحیح به سؤال j توسط آزمودنی‌ای است که در طبقه مهارتی $\boldsymbol{\alpha}_{lj}^*$ قرار دارد؛

$\bar{P}_j$ میانگین وزنی احتمال پاسخ صحیح سؤال j است که مقدار آن مطابق فرمول شماره (۳) به‌صورت زیر محاسبه می‌گردد:

$$\bar{P}_j = \sum_{l=1}^{2^{k_j^*}} \pi^{(t)}(\boldsymbol{\alpha}_{lj}^*|\boldsymbol{y}^{(t)}) \, (P(Y_j = 1|\boldsymbol{\alpha}_{lj}^*) \quad (۳)$$

$\pi^{(t)}(\boldsymbol{\alpha}_{lj}^*|\boldsymbol{y}^{(t)})$ احتمال پسین آزمودنی در طبقه مهارتی $\boldsymbol{\alpha}_{lj}^*$ پس از دریافت پاسخ به t سؤال (یعنی در پایان چرخه t است)؛

$\boldsymbol{y}^{(t)}$ مجموعه پاسخ‌های آزمودنی به سؤال‌های ارائه‌شده تا مرحله t است.

در نهایت، سؤالی که مقدار ς_j^2 آن بیشینه باشد، برای ارائه در چرخه بعدی یعنی t+1 انتخاب می‌شود.

کاپلان و همکاران (۲۰۱۵) نشان دادند که به حداکثر رساندن شاخص GDI به‌ویژه در آزمون‌های کوتاه عملکرد بسیار خوبی دارد و با به‌کارگیری بردار صفات کاهش‌یافته، محاسبات آن از نظر زمانی بهینه‌تر می‌شود.

در سال‌های اخیر، نظریه‌ها و الگوریتم‌های متعددی در ارتباط با برنامه‌های CD-CAT توسعه یافته‌اند (چنگ، ۲۰۰۹؛ کاپلان و همکاران، ۲۰۱۵؛ مک گلهن و چانگ[15]، ۲۰۰۸؛ وانگ[16]، ۲۰۱۳؛ تاتسوئوکا[17]، ۲۰۰۲؛ تاتسوکا و فرگوسن[18]، ۲۰۰۳؛ زو و همکاران، ۲۰۰۳؛ ژنگ و چانگ، ۲۰۱۶). در مطالعات

[1] Modified Posterior-Weighted Kullback-Leibler (MPWKL) Method
[2] Generalized Model Discrimination Index (GDI) Method
[3] Kaplan
[4] methods based on the CDM discrimination index
[5] Zheng & Chang
[6] Methods for attribute balancing
[7] Wang
[8] Attribute balancing design method
[9] Stratified-design method
[10] Blocked-design method
[11] Kaplan & de la Torre
[12] Binary searching design method
[13] Zheng & Wang
[14] Jensen-Shannon Divergence (JSD)
[15] McGlohen & Chang
[16] Wang
[17] Tatsuoka
[18] Tatsuoka & Ferguson

CD-CAT، الگوریتم‌های انتخاب سؤال عمدتاً مبتنی بر آنتروپی شانون[1] (SHE) هستندکه توسط تاتسوئوکا (۲۰۰۲) و تاتسوکا و فرگوسن (۲۰۰۳) معرفی شدند، و نیز بر پایه‌ی اطلاعات کولبک – لایبلر[2] (kl) که توسط زو و همکاران (۲۰۰۳) ارائه گردیده است.

با این‌حال، چنگ (۲۰۰۹) روش کولبک–لایبلر با وزن پسین[3] (PWKL) و روش اطلاعات هیبریدی کولبک–لایبلر[4] (HKL) را به‌کار گرفت. در همین راستا، وانگ (۲۰۱۳) از روش اطلاعات متقابل[5] (MI) و کاپلان و همکاران (۲۰۱۵) از روش اصلاح شده کولبک–لایبلر با وزن پسین[6] (MPWKL) و شاخص تمیز مدل تعمیم‌یافته[7] (GDI) بهره بردند. علاوه بر این، ژنگ و چانگ (۲۰۱۶) شاخص تمیز شناختی با وزن پسین[8] (PWCDI) و شاخص تمیز شناختی در سطح صفت با وزن پسین[9] (PWACDI) را توسعه دادند. همچنین مینچن و دلاتوره[10] (۲۰۱۶) شاخص واگرایی جنسن–شانون (JSD) را معرفی کردند.

یکی از جنبه‌های کلیدی در CD-CAT، الگوریتم‌های انتخاب سؤال است (چنگ، ۲۰۰۹؛ ژنگ، ۲۰۱۵؛ ژنگ و چانگ، ۲۰۱۶). در سیستم‌های CAT، انواع مختلفی از الگوریتم‌های انتخاب سؤال طراحی شده‌اند تا به نیازهای گوناگون پاسخ دهند. این الگوریتم‌ها عمدتاً بر پایه‌ی نظریه سؤال پاسخ (IRT) در مطالعات سنتی CAT بنا شده‌اند. با این‌حال، تحقیقات در زمینه CD-CAT محدودتر بوده، زیرا این حوزه به‌تازگی توجه پژوهشگران را به‌خود جلب کرده است. عوامل متعددی همچون تعداد صفات، ساختار ماتریس کیو، کیفیت سؤال‌ها، قوانین توقف و روش‌های برآورد می‌توانند بر دقت و کارایی نتایج اثرگذار باشند. ژنگ (۲۰۱۵) تأکید می‌کند که هدف اصلی الگوریتم‌های انتخاب سؤال، دستیابی به حداکثر دقت اندازه‌گیری است که این هدف در CD-CAT نیز کاملاً حائز اهمیت است.

در سال‌های اخیر، اگرچه الگوریتم‌های متعددی برای CD-CAT طراحی شده‌اند، اما اکثر آن‌ها در شرایط یکسان و قابل‌مقایسه ارزیابی نشده‌اند. در همین زمینه عشیرت و اومور–سوبول (۲۰۲۴) در پژوهشی، عملکرد الگوریتم‌های مختلف انتخاب سؤال را در CD-CAT مورد بررسی قرار دادند. این مطالعه به مقایسه الگوریتم‌ها در شرایط مختلف از جمله طول آزمون، کیفیت سؤال‌ها، تعداد صفات، و قانون توقف می‌پردازد. الگوریتم‌های مورد بررسی شامل GDI، HKL، JSD، KL، MPWKL، MI، PWACDI، PWCDI، PWKL و SHE می‌باشند. در این پژوهش، با تنظیم متغیرهای ذکرشده در سطوح مختلف، تلاش شده است تا الگوریتم‌هایی که دارای نرخ بازیابی الگوی مطلوب، زمان محاسبه کمتر و طول آزمون مناسب‌تر هستند، شناسایی شوند. نتایج نشان دادند که JSD و MPWKL دارای بالاترین نرخ بازیابی الگوها هستند، در حالی که الگوریتم‌های KL و انتخاب تصادفی، عملکرد ضعیف‌تری داشتند. افزایش کیفیت سؤال‌ها و طول آزمون منجر به بهبود نرخ بازیابی الگوها گردید. همچنین، الگوریتم‌های MI و GDI در شرایط آزمون‌های با طول ثابت، نسبت‌به SHE عملکرد بهتری از خود نشان دادند. از سوی دیگر، افزایش تعداد صفات شناختی به طولانی‌تر شدن فرایند آزمون منجر شد، در حالی که به‌کارگیری پرسش‌های با کیفیت بالا باعث کاهش طول آزمون گردید. در این میان، JSD کوتاه‌ترین طول آزمون را داشت. افزایش تعداد صفات با کاهش نرخ بازیابی الگوها و افزایش زمان محاسبه همراه بود. در این خصوص، الگوریتم GDI کمترین زمان محاسبه و MPWKL بیشترین زمان محاسبه را به خود اختصاص دادند. در نهایت، با افزایش کیفیت پرسش‌ها و طول آزمون، عملکرد الگوریتم‌ها به‌هم نزدیک‌تر شد. این یافته‌ها می‌توانند نقش به‌سزایی در بهینه‌سازی آموزش و طراحی آزمون‌های تشخیصی شناختی ایفا کنند.

۴- قانون نمره دهی (برآورد نیمرخ فرد)

خروجی اصلی سنجش انطباقی کامپیوتری تشخیصی شناختی(CD-CAT) برآورد نیمرخ صفت هر آزمون‌شونده (یعنی $\widehat{\alpha}_l$) است. در این رابطه،

i: نمایانگر آزمون‌شونده (فرد پاسخ‌گو)؛

α_i: بردار دودویی که وضعیت تسلط آزمون‌شونده i بر صفات k را نشان می‌دهد. مثلاً برای سه مهارت $\alpha_i = (1,0,1,1)$؛

k: تعداد کل صفات یا مهارت‌های شناختی مورد سنجش؛

l: شماره‌گذاری حالت‌های ممکن تسلط (الگوهای مختلف)، که برای K مهارت، تعداد این حالت‌ها 2^k خواهد بود.

برای برآورد (α_i) نیمرخ مهارت در مراحل اولیه و میانی، می‌توان از روش‌های بیشینه درست‌نمایی[11] (MLE)، بیشینه پسین[12] (MAP) یا میانگین پسین مورد انتظار[13] (EAP) استفاده کرد. همه این روش‌ها بر پایه تابع درست‌نمایی[14] $L(Y_i|\alpha_l)$ تعریف می‌شوند. این تابع مطابق فرمول شماره (۴) به‌صورت زیر بیان می‌گردد:

[1] Shannon Entropy (SHE)
[2] Kullback-Leibler (KL)
[3] Posterior-weighted Kullback–Leibler information (PWKL)
[4] Hybrid Kullback-Leibler information (HKL)
[5] Mutual information (MI) method
[6] Modified Posterior-Weighted Kullback-Leibler (MPWKL) Method
[7] Generalized Model Discrimination Index (GDI) Method
[8] Post-Weighted Cognitive Discrimination Index (PWCDI)
[9] Post-Weighted Attribute-level Cognitive Discrimination Index (PWACDI)
[10] Minchen and de la Torre
[11] Maximum Likelihood Estimate
[12] Maximum A Posteriori
[13] Expected A Posteriori
[14] Likelihood Function

$$L\left(Y_i|\alpha_l\right)=\prod_{j=1}^{J(t)}P(Y_j=1|\alpha_l)^{Y_{ij}}[1-P\left(Y_j=1\middle|\alpha_l\right)]^{1-Y_{ij}} \quad (۴)$$

در این رابطه:

Y_{ij} : پاسخ آزمون‌شونده i به سؤال j (۱ برای پاسخ صحیح، ۰ برای نادرست)؛

$P\left(Y_j=1\middle|\alpha_l\right)$: احتمال پاسخ صحیح به سؤال j در صورت داشتن الگوی تسلط α_l؛

$J(t)$: تعداد سؤال‌هایی که آزمون‌شونده تا مرحله‌ی t به آن‌ها پاسخ داده است.

از این تابع درست‌نمایی، و با در نظر گرفتن توزیع پیشین صفات $p(\alpha_c)$ توزیع پسین[1] طبقه‌های مکنون مطابق فرمول شماره (۵) به‌صورت زیر برآورد می‌شود:

$$P(\alpha_l|Y_i)=L(Y_i|\alpha_l)p(\alpha_l)/\sum_{c=1}^{2^K}L(Y_i|\alpha_c)p(\alpha_c) \quad (۵)$$

سپس، برآوردگرهای بیشینه درست‌نمایی(MLE) ، بیشینه پسین (MAP) و میانگین پسین مورد انتظار (EAP) به‌ترتیب در فرمول‌های شماره (۶)، (۷) و (۸) تعریف شده‌اند:

برآوردگر MLE: $ML(\alpha_i)=argmax_{\alpha_l}[L(Y_i|\alpha_l)]$ (۶)

برآوردگر MAP: $MAP(\alpha_i)=argmax_{\alpha_l}[P(\alpha_l|Y_i)]$ (۷)

برآوردگر EAP (میانگین پسین): $EAP(\alpha_i)=\{P(\alpha_{iK}|Y_i)\}\sum_{l=1}^{L}P(\alpha_l|Y_i)\alpha_{lK}$ (۸)

که در آن:

α_{iK} : احتمال اینکه فرد i در مهارت k تسلط داشته باشد.

α_{lK} : مقدار kام از بردار تسلط الگوی l.

برآوردگر EAP، احتمال پسینی تسلط آزمودنی i را بر هر یک از صفات k ارائه می‌دهد. این احتمال معمولاً با یک نقطه برش استاندارد مانند ۰/۵ تفسیر می‌شود (سورل[2] و همکاران، ۲۰۲۱).

هوبنر و وانگ[3] (۲۰۱۱) سه روش MLE، MAP و EAP را با یکدیگر مقایسه کرده و میزان توافق آن‌ها را در طبقه‌بندی آزمون‌شونده‌ها مورد بررسی قرار دادند. نتایج نشان داد که عملکرد روش‌های حداکثر درست‌نمایی (MLE) و حداکثر پسین (MAP) تقریباً مشابه است. با این‌حال، در طبقه‌بندی کلی صفت‌های فردی، روش میانگین پسین (EAP) دقت بالاتری نسبت‌به سایر روش‌ها دارد. در آزمون‌هایی که سطح تشخیص‌پذیری آن‌ها پایین است (مانند زمانی که پارامترهای لغزش و حدس بالا باشند)، ممکن است تفاوت‌های معناداری میان برآوردهای به‌دست‌آمده از EAP با روش‌های MLE و MAP مشاهده شود.

CD-CAT به‌عنوان روشی متمایز ازCAT، هر آزمون‌شونده را به الگویی از تسلط نسبت می‌دهد که بیشترین هم‌خوانی را با پاسخ‌های او در مجموعه سؤال‌های آزمون دارد. در این فرآیند، با بهره‌گیری از روش‌های درست‌نمایی (MLE) یا بیزی (MAP یا EAP)، احتمال تعلق فرد به هر یک از طبقه‌های مکنون برآورد می‌شود. این طبقه‌ها با الگوهای منحصربه‌فردی از تسلط بر مهارت‌ها تعریف می‌شوند و در نهایت، آزمون‌شونده به طبقه‌ای تخصیص می‌یابد که بیشترین احتمال تعلق را دارد (یو و همکاران، ۲۰۱۹).

تفاوت اصلی بین برآورد طبقه مکنون در مدل‌های تشخیصی شناختی با آزمون‌های طول ثابت و CD-CAT این است که در CD-CAT، پس از پاسخ به هر سؤال، احتمال عضویت فرد در طبقات مختلف به‌روزرسانی می‌شود و این احتمالات به‌منظور انتخاب سؤال بعدی استفاده می‌شوند؛ فرآیندی مشابه آنچه در CAT رخ می‌دهد. در حالی که در آزمون‌های ثابت، این برآورد تنها یک‌بار و پس از پاسخ‌دهی به تمامی سؤال‌ها صورت می‌گیرد.

در مدل‌های تشخیصی شناختی (CDMs) ازجمله CD-CAT، انواع متعددی از مدل‌ها وجود دارد که از نظر تابع پاسخ و نحوه استفاده از ماتریس کیو با یکدیگر تفاوت دارند. به‌عنوان نمونه، مدل DINA یک مدل پیوندی[4] محسوب می‌شود، چراکه برای کسب احتمال بالای پاسخ صحیح، تسلط کامل بر تمامی مهارت‌های موردنیاز ضروری است. در مقابل، مدل‌هایی که صرفاً تسلط بر بخشی از مهارت‌ها را کافی می‌دانند، به‌عنوان مدل‌های منفصل[5] شناخته می‌شوند. در هر یک از این رویکردها، مدل‌های جایگزین دیگری نیز وجود دارند که از نظر تابع پاسخ، پارامترهای سؤال و سایر ویژگی‌ها با مدل

[1] Posterior Distribution
[2] Sorrel
[3] Huebner & Wang
[4] Conjunctive
[5] Disjunctive

DINA متفاوت‌اند. با این حال، تمامی این مدل‌ها امکان برآورد طبقه تسلط مکنون را با استفاده از روش‌های حداکثر درست‌نمایی یا بیزی فراهم می‌سازند (ویس و شاهین، ۲۰۲۴).

۵- قاعده توقف

در CD-CAT پیشنهادهای مختلفی برای تعیین زمان پایان آزمون ارائه شده‌اند، اما تحقیقات اندکی به‌صورت مستقیم به این موضوع پرداخته‌اند. یکی از روش‌های شناخته‌شده، روش طول ثابت است که در برخی سیستم‌های CAT به‌کار رفته (چنگ، ۲۰۰۹؛ وانگ و همکاران، ۲۰۱۱؛ زو و همکاران، ۲۰۰۳). در این روش، سؤال‌ها به‌صورت انطباقی و به‌ترتیب یک‌به‌یک انتخاب می‌شوند تا زمانی که تعداد مشخصی از آن‌ها ارائه گردد. همان‌طور که در CAT نیز مشاهده شده، این روش تأثیر مستقیمی بر کارایی آزمون ندارد، زیرا تلاشی برای به حداقل رساندن تعداد سؤال‌ها برای هر آزمون‌دهنده صورت نمی‌گیرد. در نتیجه، اجرای CD-CAT با این روش ممکن است منجربه طبقه‌بندی‌هایی با دقت‌های متفاوت شود، که به نوبه خود، روایی آزمون برای آزمودنی‌های مختلف را تحت تأثیر قرار می‌دهد. علت این تفاوت آن است که برخی آزمودنی‌ها برای طبقه‌بندی دقیق‌تر به سؤال‌های بیشتری نیاز دارند، در حالی که برخی دیگر با زیرمجموعه‌ای کوچک‌تر از سؤال‌های مناسب، به‌خوبی طبقه‌بندی می‌شوند.

به‌طور مشخص، همانند CAT، میزان سودمندی یک معیار توقف در CD-CAT نیز بستگی‌به کیفیت بانک سؤال و تعامل آن با مدل شناختی مورد استفاده دارد. یک برنامه CD-CAT می‌تواند دارای طول ثابت یا متغیر باشد. در حالت طول ثابت، آزمون زمانی پایان می‌یابد که تعداد مشخصی از سؤال‌ها به هر آزمودنی ارائه شود. در مقابل، برنامه‌هایی با طول متغیر از قوانین متنوع‌تری برای خاتمه آزمون بهره می‌برند (وون، داویر و لی، ۲۰۱۹). اکثر مطالعات پیشین از قانون طول ثابت استفاده کرده‌اند. با این حال، CD-CAT که نسخه انطباقی کامپیوتری سنجش تشخیصی شناختی (CDA) به‌شمار می‌رود، نیازمند یک قانون انعطاف‌پذیر برای خاتمه آزمون است، به‌طوری که بتواند در سطح مناسبی از اطمینان، فرایند را متوقف کند (گوا و ژنگ [1]، ۲۰۱۹).

در آزمون‌های CD-CAT با طول متغیر، قوانین توقف متعددی مطرح شده‌اند. به‌عنوان مثال، تاتسوئوکا (۲۰۰۲) از قانونی استفاده کرد که آزمون را زمانی متوقف می‌کرد که احتمال پسین تعلق آزمودنی به یک الگوی چیرگی صفات به ۰/۸ برسد. هسو[2] و همکاران (۲۰۱۳) نیز در این حوزه پژوهش‌هایی انجام دادند و دو قانون توقف برای CD-CAT با طول متغیر پیشنهاد کردند:

قانون اول: آزمون زمانی متوقف می‌شود که بیشترین احتمال پسین (در میان تمام حالت‌های مکنون ممکن) کمتر از یک مقدار مشخص، مثلاً ۰/۷ نباشد.

قانون دوم: افزون بر شرط اول، دومین احتمال پسین نیز باید کمتر یا مساوی یک مقدار از پیش تعیین‌شده مانند ۰/۱ باشد (وون داویر و لی، ۲۰۱۹).

در CD-CAT، چند روش برای پایان آزمون با طول متغیر پیشنهاد شده‌اند. یکی از این روش‌ها این است که آزمون زمانی خاتمه یابد که احتمال تعلق آزمودنی به هر یک از طبقات، برابر یا بیشتر از مقدار مشخصی مانند ۰/۸ باشد (تاتسوکا، ۲۰۰۲). از آنجا که این مقدار به‌صورت دلخواه تعیین می‌شود، امکان تغییر آن وجود دارد؛ برای مثال، هسو و همکارانش در سال ۲۰۱۳ مقدار ۰/۷ را پیشنهاد کردند. بدیهی است که افزایش این مقدار احتمالاً به افزایش طول آزمون منجر خواهد شد، در حالی که کاهش آن باعث کوتاه‌تر شدن آزمون‌ها می‌شود. این رویکرد، مشابه قانون توقف بر اساس خطای استاندارد ثابت در آزمون‌های CAT است. در CD-CAT، احتمال بالا برای یک طبقه نشان‌دهنده توزیع قله‌دارتر است، در حالی که احتمال پایین‌تر بیانگر توزیع هموارتر در میان طبقه‌های مکنون است.

در همین راستا، تاتسوئوکا (۲۰۰۲) اصلاحیه‌ای برای قانون توقف پیشنهاد داد که بعداً توسط هسو و همکاران (۲۰۱۳) به‌تفصیل بررسی شد. این اصلاحیه شامل افزودن یک مقدار دوم اختیاری به قانون توقف بود، به‌طوری که نه تنها بالاترین احتمال پسین در نظر گرفته می‌شد، بلکه دومین احتمال نیز باید کمتر یا مساوی مقدار دوم مشخصی (مثلاً ۰/۱) باشد. به باور آن‌ها، این معیار دوگانه می‌تواند توزیع احتمال گسسته بین طبقات را بهتر توصیف کند و نسبت به توزیعی که تنها از یک معیار استفاده می‌کند، نمای قله‌دارتر و شفاف‌تری ارائه دهد (ویس و شاهین، ۲۰۲۴). انتظار می‌رود در آینده، قوانین توقف دیگری نیز ارائه شده و در تحقیقات CD-CAT با قوانین موجود مقایسه شوند.

مطالعات مختلف نشان داده‌اند که رویکرد CD-CAT از نظر کارایی و روایی، عملکرد قابل قبولی ارائه می‌دهد، و این امر نشان‌دهنده پتانسیل آن برای به‌کارگیری در ارزیابی‌های گسترده است. با این حال، به‌دلیل پیچیدگی بالای الگوریتم‌ها و فرآیند طراحی سیستم، اغلب مطالعات موجود از تنظیمات نسبتاً ثابتی در مؤلفه‌های کلیدی CD-CAT، مانند مدل‌سازی، انتخاب سؤال و قواعد خاتمه بهره برده‌اند (لی و همکاران، ۲۰۲۳).

علیرغم این پیشرفت‌ها، هنوز تا ادغام کامل این نوآوری‌ها در آزمون‌های رسمی و راهبردهای آموزشی، فاصله وجود دارد (راوند و بقایی، ۲۰۲۰). همچنین، شایان ذکر است که تا زمان انجام این پژوهش، هیچ مطالعه‌ای در ایران پیرامون ساخت آزمونی بر اساس روش سنجش انطباقی کامپیوتری

[1] Guo & Zheng

[2] Hsu

تشخیصی شناختی(CD-CAT) صورت نگرفته است. این حوزه در ایران تاکنون مورد توجه محدود قرار گرفته و پژوهشگران داخلی هنوز در مسیر طراحی آزمون براساس این رویکرد گام برنداشته‌اند. امید است این پژوهش بتواند به توسعه علمی و تولید ابزارهای نوین ارزیابی در کشور کمک کند.

نرم‌افزارها و بسترهای اجرایی درCD-CAT

برای پیاده‌سازی مؤثر مدل‌های تشخیصی شناختی در چارچوب CD-CAT، مجموعه‌ای از ابزارهای آماری و نرم‌افزاری در دسترس پژوهشگران قرار دارد که هر یک، متناسب با هدف تحلیل، سطح تخصص کاربر، و نوع داده‌های مورد استفاده، قابلیت‌های منحصربه‌فردی ارائه می‌کنند. در این میان، بسته‌ی قدرتمند و چندمنظوره‌ی GDINA در محیط نرم افزاری R یکی از برجسته‌ترین ابزارهای تحلیل در حوزه‌ی CD-CAT به‌شمار می‌رود. این بسته امکاناتی مانند مدل‌سازی ماتریس کیو، تخمین دقیق مدل‌های DINA، DINO و G-DINA، اجرای شبیه‌سازی آزمون‌های انطباقی، و تحلیل نیم‌رخ‌های چیرگی را در اختیار کاربران قرار می‌دهد (ما و دلاتوره[1]، ۲۰۲۰). GDINA علاوه‌بر دقت بالا و انعطاف‌پذیری در تحلیل داده‌های واقعی، به پژوهشگران امکان می‌دهد ساختار شناختی آزمون را متناسب با نیازهای پژوهشی یا آموزشی طراحی و پیاده‌سازی کنند.

در کنار ابزارهای تحلیلی، بسترهای اجرایی نیز نقش کلیدی در موفقیت پیاده‌سازی آزمون‌های CD-CAT ایفا می‌کنند. در این زمینه، پلتفرم Concerto، که یک سامانه‌ی منبع‌باز[2] برای طراحی و اجرای آزمون‌های انطباقی مبتنی بر وب است و بر پایه زبان‌های برنامه‌نویسی R و PHP که مخفف Hypertext Preprocessor، زبان برنامه‌نویسی سمت سرور است توسعه یافته، امکان طراحی، اجرا و مدیریت آزمون‌های تعاملی، تطبیقی و شناخت‌محور را بدون نیاز به کدنویسی پیچیده فراهم می‌سازد.

طراحی ماژولار و محیط کاربرپسند Concerto، آن‌را به انتخابی مناسب برای پژوهشگران حوزه روانسنجی، به‌ویژه کسانی‌که در پی اجرای آزمون‌های پیشرفته در مقیاس وسیع هستند، تبدیل کرده است. این سامانه در حال حاضر توسط دانشگاه کمبریج به‌عنوان یکی از ابزارهای پیشنهادی در زمینه آزمون‌های تطبیقی معرفی شده است.

همچنین، نرم‌افزار پیشرفته مدل‌سازی انعطاف‌پذیر نظریه سؤال پاسخ چندسطحی[3] (FlexMIRT) برای تحلیل‌های پیچیده در حوزه اندازه‌گیری چندبعدی، برآورد صفات پیوسته و گسسته، و طراحی آزمون‌های تشخیصی دقیق مورد استفاده قرار می‌گیرد. بهره‌گیری از این ابزار، گامی اساسی در جهت کاربردی‌سازی مفاهیم نظری CD-CAT در محیط‌های آموزشی و پژوهشی محسوب می‌شود (کای[4]، ۲۰۱۷).

بحث و نتیجه‌گیری

در این مقاله، مروری جامع و نظام‌مند بر اجزای کلیدی رویکرد سنجش انطباقی کامپیوتری تشخیصی‌شناختی (CD-CAT) ارائه شد؛ به‌ویژه، مؤلفه‌هایی مانند معیارهای توقف، الگوریتم‌های انتخاب سؤال، و مدل‌های تشخیصی و ساختار ماتریس کیو مورد بررسی قرار گرفتند. تحلیل ادبیات پژوهش نشان داد که با وجود تمرکز مطالعات متعدد بر سیستم‌های انتخاب سؤال (ISS)، مؤلفه‌هایی نظیر قواعد توقف، ساختار ماتریس کیو و برآورد همزمان مهارت و توانایی، کمتر مورد توجه قرار گرفته‌اند. این خلأ پژوهشی، زمینه‌ای مهم برای مطالعات آینده فراهم می‌کند و نشان‌دهنده‌ی نیاز به توسعه‌ی نظام‌مندتر این حوزه است. این عدم‌توازن پژوهشی می‌تواند مانعی جدی در بهره‌برداری بهینه از ظرفیت‌های CD-CAT باشد.

مقایسه الگوریتم‌های انتخاب سؤال نظیرGDI، JSD، MPWKL، SHE، KL و CDI نشان داد که انتخاب آگاهانه شاخص مناسب بسته به هدف آزمون (مثلاً دقت، سرعت، یا بازخورد شناختی)، تأثیری تعیین‌کننده بر دقت طبقه‌بندی، طول آزمون، و کیفیت ارزیابی شناختی دارد. الگوریتم‌هایی مانند MPWKL و PWKL به‌ویژه در مراحل آغازین آزمون عملکرد بهتری دارند، در حالی که شاخص‌هایی مانند GDI و CDI برای مدل‌های پیچیده‌تر مناسب‌ترند. در کنار تحلیل‌های مفهومی، معرفی نرم‌افزارها و بسترهای اجرایی مانند بسته‌ی قدرتمند GDINA در محیط R، ابزارهایی مانند CDM و mirt، و نیز سامانه‌های اجرایی مانند Concerto و FlexMIRT، چشم‌انداز عملیاتی‌سازی CD-CAT را برای پژوهشگران روشن‌تر کرد. بهره‌گیری از این ابزارها، انتقال این رویکرد از سطح نظری به سطح اجرا را تسهیل کرده و مسیر توسعه آزمون‌های انطباقی شناخت‌محور را هموارتر ساخته است.

CD-CAT در مقایسه با آزمون‌های سنتی مبتنی بر IRT یا آزمون‌های مداد–کاغذی، امکان سنجش چندبعدی، دقیق و فردمحور را فراهم می‌سازد. برخلاف آزمون‌های سنتی مانند CAT که تنها به برآورد یک‌بعدی از توانایی کلی آزمودنی بسنده می‌کنند، رویکرد CD-CAT با بهره‌گیری از مدل‌های تشخیصی شناختی و ساختار ماتریس‌کیو، امکان استخراج نیم‌رخ‌های چندبعدی از مهارت‌های شناختی افراد را فراهم می‌سازد (چانگ[5]، ۲۰۱۴).

با توجه به این ظرفیت‌ها، انتظار می‌رود CD-CAT در آینده‌ای نزدیک در حوزه‌های آموزش، روان‌شناسی، سلامت روان و تشخیص‌های بالینی جایگاهی محوری بیابد. آشنایی متخصصان با اصول طراحی، تحلیل و پیاده‌سازی این رویکرد، گامی بنیادین در جهت ارتقای کیفیت سنجش‌های

[1] Ma & de la Torre
[2] Open Source
[3] Flexible Multilevel Item Response Theory Modeling Software (FlexMIRT)
[4] Cai
[5] Chang

انسان‌محور و داده‌محور است. شواهد مرور نظام‌مند اخیر نیز نشان می‌دهد که CD-CAT در حال تبدیل‌شدن به یکی از رویکردهای پیشرو در آزمون‌سازی هوشمند و داده‌محور است و انتظار می‌رود سهم آن در ارزیابی‌های آموزشی و تشخیصی در سال‌های آینده افزایش یابد.

در نهایت، پیشنهاد می‌شود که مطالعات آینده با تمرکز بیشتر بر مؤلفه‌های کمتر کاوش‌شده مانند قواعد توقف، ساختار ماتریس کیو، و تخمین‌های هم‌زمان θ و α انجام گیرند. همچنین، پژوهش‌های بومی‌سازی‌شده در بافت فرهنگی ایران، می‌تواند به طراحی آزمون‌هایی منجر شود که نه‌تنها از حیث روانسنجی معتبرند، بلکه به نیازهای محلی نیز پاسخ می‌دهند.

سپاس‌گزاری

نویسندگان از کسانی که مطالب مفیدی درباره محتوای مطالعه حاضر پیشنهاد دادند تشکر می‌کنند.

ملاحظات اخلاقی

این مطالعه مبتنی‌بر مرور ادبیات بوده و شامل شرکت‌کنندگان انسانی نمی‌باشد. با وجود موارد اخلاقی از جمله امانت داری و ذکر منابع درست رعایت شده است.

حامی مالی

هزینه‌های انجام این پژوهش و نگارش مقاله به‌طور کامل توسط نویسندگان تأمین شده و تحقیق حاضر از هیچ‌گونه حمایت مالی بیرونی بهره‌مند نبوده است.

مشارکت نویسندگان

این مقاله برگرفته از رساله دکتری است. نویسنده اول دانشجوی دکتری است و نویسندگان دوم و سوم به‌عنوان استادان راهنما و نویسندگان چهارم و پنجم به عنوان مشاوران علمی پژوهش در این رساله مشارکت داشته‌اند.

تعارض منافع

نویسندگان اعلام می‌دارند که هیچ گونه تعارض منافعی ندارند.

منابع

مینائی، اصغر(۱۳۹۱). *مدل پردازی تشخیصی شناختی (CDM) سؤال‌های ریاضیات تیمز ۲۰۰۷ در دانش آموزان پایه هشتم ایران با استفاده از مدل یکپارچه با پارامترپردازی مجدد (RUM) و مقایسه مهارت‌های ریاضی دانش‌آموزان دختر و پسر*. رساله‌ی دکتری. دانشگاه علامه طباطبائی.

References

Andrade Medeiros, A. M., & Muniz, C. A. (2022). Mathematical learning difficulties: a subjective production. *The Mathematics Enthusiast*, *19(1)*, 28-54.

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.

Bi, H., Ma, H., Huang, Z., Yin, Y., Liu, Q., Chen, E., Su, Y., & Wang, S. (2020). *Quality meets diversity: A model-agnostic framework for computerized adaptive testing*. 2020 IEEE International Conference on Data Mining (ICDM).

Boussakuk, M., Bouchboua, A., El Ghazi, M., El Bekkali, M., & Fattah, M. (2021). Design of Computerized Adaptive Testing Module into Our Dynamic Adaptive Hypermedia System. *International Journal of Emerging Technologies in Learning (iJET)*, *16(18)*, 113-128.

Cai, L. (2017). *flexMIRT® (Version 3.51)* [Computer software]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Chang, h.-h. (2014). Psychometrics Behind Computerized Adaptive Testing. *Psychometrika*, *80*.

Chen, P., Xin, T., Wang, C., & Chang, H.-H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika*, *77(2)*, 201-222.

Cheng, Y. (2009). When Cognitive Diagnosis Meets Computerized Adaptive Testing: CD-CAT. *Psychometrika*, *74(4)*, 619-632.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

De La Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of educational and behavioral statistics*, *34(1)*, 115-130.

de la Torre, J., & Douglas, J. A. (2004). Higher-Order Latent Trait Models for Cognitive Diagnosis. *Psychometrika*, *69*(3), 333–353.

de la Torre, J. (2011). The Generalized DINA Model Framework. *Psychometrika*, *76*(2), 179–199.

Guo, L., & Zheng, C. (2019). Termination Rules for Variable-Length CD-CAT From the Information Theory Perspective. *Frontiers in Psychology*, *10*, 1122.

Huebner, A. (2019). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research, and Evaluation*, *15*(1), 3.

Huebner, A., & Wang, C. (2011). A Note on Comparing Examinee Classification Methods for Cognitive Diagnosis Models. *Educational and Psychological Measurement, 71(2)*, 407-419.

Jafri, S. S. M. (2007). Computerized adaptive testing using neural networks. *A thesis presented to the deanship of graduate studies in partial fulfillment of the requirements for the degree Master of Science. King Fahd University of Petroleum & Minerals Dhahran, Eastern Province, Saudi Arabia*.

Javidanmehr, Z., & Anani Sarab, M. R. (2017). An investigation of the prevalence and difficulty of reading comprehension's sub-skills by the G-DINA model. *Critical Language and Literary studies*, *14*(19), 99-118.

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology, 33(4)*, 410-422.

Kaplan, M., & de la Torre, J. (2020). A Blocked-CAT Procedure for CD-CAT. *Applied Psychological Measurement*, *44(1),* 49-64.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New Item Selection Methods for Cognitive Diagnosis Computerized Adaptive Testing. *Appl Psychol Meas*, *39*(3), 167-188.

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.*

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*.

Liu, H.-Y., You, X.-F., Wang, W.-Y., Ding, S.-L., & Chang, H.-H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification, 30(2), 152-172.*

Liu, Q., Zhuang, Y., Bi, H., Huang, Z., Huang, W., Li, J., Yu, J., Liu, Z., Hu, Z., & Hong, Y. (2024). *Survey of computerized adaptive testing: A machine learning perspective*. arXiv

Liu, Y., Zhang, T., Wang, X., Yu, G., & Li, T. (2023). New development of cognitive diagnosis models. *Frontiers of Computer Science*, *17(1)*, *171604.*

Luo, F., Wang, X., Cai, Y., & Tu, D. (2023). Two efficient selection methods for high-dimensional CD-CAT utilizing max-marginals factor from MAP query and ensemble learning approach. *British Journal of Mathematical and Statistical Psychology*, *76(2), 283-311.*

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93*, 1–26.

Magis, D., Yan, D., & Von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. Springer

McClelland, T., & Cuevas, J. (2020). A comparison of computer-based testing and paper and pencil testing in mathematics assessment. *The Online Journal of New Horizons in Education*, *10*(2), *78-89.*

McGlohen, M., & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, *40(3), 808-821.*

Minaei, A. (2012). *Modeling the cognitive diagnostic assessment (CDM) of TIMSS 2007 mathematics items among Iranian eighth grade students using the reparameterized unified model (RUM) and comparing math skills of male and female students* (Doctoral dissertation, Allameh Tabataba'i University). [Persion]

Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing*, *20(1)*, *24-56.*

Riley, B. B., Conrad, K. J., Bezruczko, N., & Dennis, M. L. (2007). Relative precision, efficiency, and construct validity of different starting and stopping rules for a computerized adaptive test: the GAIN substance problem scale. *Journal of applied measurement, 8(1), 48-64.*

Rudner, L. M., (1998). Item Banking. *Practical Assessment, Research, and Evaluation* ,*6(1)*: 4.

Rupp, A. A., & Templin, J. (2008). The Effects of Q-Matrix Misspecification on Parameter Estimates and Classification Accuracy in the DINA Model. *Educational and Psychological Measurement, 68(1)*, *78-96.*

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives, 16(1)*, 1-17.

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.*

Sorrel, M. A., Nájera, P., & Abad, F. J. (2021). cdcatR: An R package for cognitive diagnostic computerized adaptive testing. *Psych, 3(3),* 386-403.

Su, K. (2022). *Implementation of a Diagnostic Classification Model for Middle-School Physics*. The University of North Carolina at Greensboro.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *51*(3), 337-350.

Tatsuoka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *65*(1), 143-157.

Templin, J., & Hoffman, L. (2013). Obtaining Diagnostic Classification Model Estimates Using Mplus. *Educational measurement: Issues and practice*, *32*(2), 37-50.

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.

von Davier, M., & Lee, Y.-S. (2019). Handbook of diagnostic classification models. *Cham: Springer International Publishing*.

von Davier, M., Cheng, Y., & Cheng, C. (2014). Multistage testing using diagnostic models. *Computerized multistage testing: Theory and applications*, 219-227.

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement, 73(6),* 1017-1035.

Wang, C. (2021). On interim cognitive diagnostic computerized adaptive testing in learning context. *Applied psychological measurement*, *45(4), 235-252.*

Wang, C., Chang, H. H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of educational measurement*, *48*(3), 255-273.

Wang, D., & Tu, D. (2021). The application of cognitive diagnostic computerized adaptive testing on diagnosis and assessment of psychological disorder. J*iangxi Normal University*, *45*, *111-117.*

Wang, S., McCall, M., Jiao, H., & Harris, G. (2013). Construct Validity and Measurement Invariance of Computerized Adaptive Testing: Application to Measures of Academic Progress (MAP) Using Confirmatory Factor Analysis. *Journal of Educational and Developmental Psychology*, *3*.

Wang, Y., Sun, X., Chong, W., & Xin, T. (2020). Attribute discrimination index-based method to balance attribute coverage for short-length cognitive diagnostic computerized adaptive testing. *Frontiers in Psychology*, *11*, *224.*

Weiss, D. J., & Sahin, A. (2024). *Computerized Adaptive Testing: From Concept to Implementation*. Guilford Publications.

Weissman, A. (2016). IRT-based multistage testing. *In Computerized multistage testing* (pp. 191-206). Chapman and Hall/CRC.

Xu, L., Jin, R., Huang, F., Zhou, Y., Li, Z., & Zhang, M. (2020). Development of Computerized Adaptive Testing for Emotion Regulation. *Frontiers in Psychology*, *11*.

Xu, X., Chang, H., & Douglas, J. (2003). A simulation study to compare CAT strategies for cognitive diagnosis. *annual meeting of the American Educational Research* Association, Chicago.

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*.

Yu, X., Cheng, Y., & Chang, H. H. (2019). Recent Developments in Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT): A Comprehensive Review. *In Methodology of Educational Measurement and Assessment* (pp. 307-331). (Methodology of Educational Measurement and Assessment). Springer.

Zheng, C. (2015). *Some practical item selection algorithms in cognitive diagnostic computerized adaptive testing smart diagnosis for smart learning*. University of Illinois at Urbana-Champaign.

Zheng, C., & Chang, H.-H. (2016). High-Efficiency Response Distribution–Based Item Selection Algorithms for Short-Length Cognitive Diagnostic Computerized Adaptive Testing. *Applied Psychological Measurement*, *40(8), 608-624*.

Zheng, C., & Wang, C. (2017). Application of Binary Searching for Item Exposure Control in Cognitive Diagnostic Computerized Adaptive Testing. *Applied Psychological Measurement, 41(7), 561-576.*

Zhu, Z., & Chang, H. (2021). Application and prospect of cognitive diagnostic computerized adaptive testing. *China Examinations*, *1*, *41-46.*