International Journal of Assessment Tools in Education
2025, Vol. 12, No. 3, 843–870
https://doi.org/10.21449/ijate.1533622
journal homepage: https://dergipark.org.tr/en/pub/ijate
Review Article

Charting the trajectory of cognitive diagnostic assessment in language assessment: A systematic literature review

Muhamad Firdaus Mohd Noh[1*], **Mohd Effendi Ewan Mohd Matore**[2], **Nur Ainil Sulaiman**[3]

[1]Universiti Kebangsaan Malaysia, Faculty of Education, Bangi, Malaysia
[2]Universiti Kebangsaan Malaysia, Faculty of Education, Research Centre of Leadership and Educational Policy, Bangi, Malaysia
[3]Universiti Kebangsaan Malaysia, Faculty of Education, Centre of Teaching and Learning Innovation, Bangi, Malaysia

ARTICLE HISTORY
Received: Aug. 15, 2024
Accepted: June 21, 2025

Keywords:
Cognitive diagnostic assessment,
Cognitive diagnostic models,
Classification diagnostic models,
Language assessment,
Systematic literature review.

Abstract: The education system has relied on conventional single-score assessments, which provide limited insights into students' cognitive abilities. Cognitive diagnostic assessment (CDA) offers a promising alternative by providing detailed information on students' cognitive processes. The review aims to explore the recent trends of cognitive diagnostic assessment publications in language assessment by examining the years of publications, geographical locations, research methodologies, focused skills, and research samples. It further investigates the implementation of CDA studies in language assessment and identifies proposed future research directions in this field. Findings show a surge in CDA studies since 2019, with China and Iran leading, predominantly focusing on receptive skills and retrofitting methods over true CDA. The review shows that CDA accentuates cognitive model design, Q-matrix construction, item development, comparison of cognitive diagnostic models (CDMs), data analysis, and mastery profile generation. Previous research has proposed recommendations related to item and instrument development, attribute selection, sample selection, and the effectiveness of CDA. The study limitations include potential biases from publication trends and the reliance on existing literature, possibly excluding unpublished works. Future efforts should focus on developing comprehensive cognitive models, balancing the assessment of receptive and productive skills, and expanding the analysis of item parameters. The review's findings imply that practitioners must align existing tests with diagnostic needs and incorporate various item parameters to enhance assessment accuracy. It also underscores the implication that educational policymakers should support refined cognitive models and that teachers must be trained to use advanced diagnostic tools for effective instructional strategies.

1. INTRODUCTION

The education system has long relied on conventional assessment methods that primarily use single-score outcomes to evaluate student performance. However, these methods fail to capture

*CONTACT: Muhamad Firdaus MOHD NOH p132669@siswa.ukm.edu.my Universiti Kebangsaan Malaysia, Faculty of Education, Bangi, Malaysia

e-ISSN: 2148-7456

the complexities of students' cognitive abilities (Huang & Bolt, 2023), hindering effective interventions and curriculum development for language acquisition (Ma *et al*., 2020). Conventional assessments cannot provide educators and policymakers with detailed insights into the specific conceptions and misconceptions of individual learners (Thompson & Clark, 2024). While they offer a snapshot of overall performance, they do not identify the specific cognitive skills and knowledge areas that need attention. Cognitive diagnostic assessment (CDA) offers a promising alternative, capable of generating detailed information about students' cognitive processes and providing a more comprehensive understanding of their learning profiles (Mohd Noh & Mohd Matore, 2024). CDA is an innovative approach to educational measurement that integrates psychological and measurement theories to diagnose students' specific cognitive strengths and weaknesses. Unlike traditional assessments that provide a single overall score, CDA uses detailed diagnostic models to break down the cognitive processes involved in answering test items, offering granular insights into which specific skills and knowledge areas a student has mastered or needs improvement in (Lee *et al*., 2025).

CDA emerged as a response to the inadequacies of traditional educational measurement practices, which often failed to capture the complexities of human cognition and learning processes. Traditional assessments, such as those focusing solely on intelligence quotient or academic skills, provided a limited view of students' cognitive capabilities (von Davier & Lee, 2019). CDA, on the other hand, offers a holistic and contextual understanding of individual intelligence and cognitive processes by combining insights from psychological theories and measurement approaches (Ketabi *et al*., 2021), including Piaget's Cognitive Development Theory, Miller's Information Processing Theory, and advanced measurement approaches like Item Response Theory (IRT). The historical development of CDA is rooted in the evolution of psychological and measurement theories. Early educational psychology focused on intelligence measurement through specific tests, emphasizing intellectual and academic skills (van der Linden, 2016).

The historical development of CDA reflects a shift from focusing solely on intelligence measurement to a broader evaluation of cognitive processes—how information is handled, stored, and processed. This shift was influenced significantly by cognitive psychology, which emphasizes complex cognitive functions over simple memorization or task completion, and IRT, which helps in understanding how item characteristics relate to test-taker abilities at different levels. This integration of psychological and measurement theories led to the development of CDA models, enabling a more comprehensive and contextual assessment of individual intelligence and cognitive processes (Sessoms & Henson, 2018). CDA emerged in the early 1980s as a response to the growing recognition of the limitations of traditional educational measurement practices and the influence of cognitive psychology on understanding human cognition and learning processes (Leighton & Gierl, 2007b). Key publications, such as Embretson's work in 1984, emphasized the importance of understanding cognitive processes, strategies, and knowledge storage involved in performance, laying the foundation for what would become CDA (Embretson, 1984). Cognitive psychologists also explored the relationship between item difficulty and cognitive processes, further contributing to the development of CDA (Fischer & Formann, 1982). By the late 1980s, the need for collaboration between educational measurement (psychometrics) and cognitive psychology to improve the design, scoring, and reporting of educational assessments was widely recognized (Wang & Lu, 2021). This integration has led to the development of sophisticated CDA models that offer insights into individual learning profiles, highlighting both the depth and context of cognitive abilities.

1.1 Cognitive Diagnostic Assessment in Language Assessment

In the context of language assessment, CDA has been employed to provide detailed insights into language learners' strengths and weaknesses. Unlike traditional psychometric approaches that predominantly focus on ranking students based on the statistical properties of test items

(Lim & Bangeranye, 2024), CDA employs a diagnostic approach, which understands language processing as a function of both verbal and visual information channels (Ma *et al*., 2020). This diagnostic approach helps in identifying specific areas of difficulty and guides the development of targeted instructional strategies. One common approach in CDA research is the modification of existing high-stakes assessments for diagnostic purposes, known as retrofitting (Mohd Noh & Mohd Matore, 2024). Retrofitting involves reanalyzing existing tests to ensure they provide relevant and useful diagnostic information (Ravand & Baghaei, 2019). This method has been applied to various assessments, including the Programme for International Student Assessment (PISA) (Chen & Chen, 2015, 2016), the Michigan English Language Assessment Battery (MELAB) (Li *et al*., 2015), the Test of English as a Foreign Language (TOEFL) (Safari & Ahmadi, 2023), and the International English Language Testing System (IELTS) (Mirzaei *et al*., 2020).

Retrofitting offers cost and time efficiencies by leveraging existing assessments, but it may face challenges such as ensuring a sufficient number of items to measure cognitive skills accurately (Gierl & Cui, 2008). Despite the prevalence of retrofitting, some researchers advocate for the development of entirely new CDA items and assessments. This approach, though more resource-intensive, allows for a more detailed and personalized analysis of learners' cognitive strengths and weaknesses (Fan *et al*., 2021). Studies such as Toprak and Cakir (2021) which developed a diagnostic reading test for university students in Turkey, and Alavi and Ranjbaran (2018), which created a diagnostic reading test in Iran, highlight the benefits of this approach. This approach, more aligned with cognitive psychological principles, enables the creation of assessments that are specifically tailored to diagnose detailed aspects of cognitive processing in language learning. By integrating these theoretical frameworks, assessments can be designed to not only evaluate language proficiency but also to understand how different cognitive skills contribute to language learning, allowing for more personalized and effective educational interventions (Ketabi *et al*., 2021).

1.2. Purpose of the Study

Over the years, several reviews have been documented by scholars to analyze the trends and patterns of studies conducted in the area of CDA. Lee and Sawaki (2009) cover the historical context and application of CDA in language assessment, emphasizing the importance of collaboration between language assessment experts and psychometricians. Then, Javidanmehr and Sarab (2017) delve into the importance of statistical models in extracting information about students' cognitive strengths and weaknesses. Sessoms and Henson (2018) provide a review of diagnostic classification models (DCMs) across various fields by analyzing existing studies, and addressing common themes, strengths, and weaknesses of DCM applications. Ravand and Baghaei (2020) review the current state of DCMs by highlighting the limited adoption of DCMs for their intended purposes, with most research focusing on methodological development and retrofitting existing tests. Fan *et al*. (2021) focus on integrating diagnostic assessment into an English as a Foreign Language (EFL) reading curriculum. Recently, Mei and Chen (2022) categorize research topics into application, optimization, and practical uses of CDA.

While the existing reviews provide an overview of CDA and DCMs, notable aspects remain underexplored, particularly in understanding how CDA studies have been distributed across different skills, geographical contexts, and methodologies. Such an understanding is crucial to assess the breadth of current research and to identify areas that require further investigation. Moreover, the distinction between retrofitting studies and true CDA studies calls for a deeper exploration of how CDA has been implemented in language assessment and the implications of each approach on diagnostic accuracy and practical use. In addition, while some previous reviews have discussed future directions (Alallo *et al*., 2023; Mei & Chen, 2022), these discussions lack the systematic thematic synthesis and methodological specificity needed to guide future CDA scholarship in a focused and impactful manner. Mei and Chen (2022) have discussed several future research directions, particularly in relation to productive skills,

diagnostic feedback, and inter-attribute relationships; however, their discussion lacks attention to the development of CDA-specific tests, Q-matrix construction, model selection, and statistical analysis. In contrast, Alallo *et al*. (2023) address important areas such as hierarchical attribute relationships, productive skills, and measurement invariance in CDMs, but the recommendations lack thematic coherence, methodological detail, and adequate consideration of Q-matrix development, item construction, and validation procedures. A more coherent structure that explicitly addresses each previously identified gap, detailed with feasible methodological pathways, would significantly strengthen its utility to researchers in the field. Addressing these gaps will not only consolidate existing knowledge but also inform future research directions in the field. Therefore, the review aims to explore the recent trends of cognitive diagnostic assessment publications in language assessment. It further investigates the methodological features of cognitive diagnostic assessment studies in language assessment and identifies proposed future research directions in this field. Specifically, this review seeks to answer the following research questions:

1. What is the recent trend of cognitive diagnostic assessment publications in language assessment in terms of years of publications, geographical locations, focused skills and research samples?
2. How do methodological approaches in cognitive diagnostic assessment studies of language assessment vary in terms of study designs and statistical analyses?
3. What future research directions have been proposed in the cognitive diagnostic assessment studies of language assessment?

These research questions aim to provide a comprehensive understanding of the current state of CDA research in language assessment. The first question maps the existing literature by identifying key publication patterns and thematic foci, which can help reveal both mature and emerging areas of study. The second question delves into the methodological dimensions of CDA research, highlighting how design choices and analytical approaches influence the quality and generalizability of findings. The third question synthesizes the future directions proposed by past studies, offering a clearer roadmap for advancing the field through theoretically grounded and methodologically sound research.

2. METHOD

The review process for the current study was detailed, covering aspects such as publication standards, database selection, identification, screening, eligibility and quality appraisal. The review adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 (PRISMA 2020) as its publication standard to improve and standardize the reporting quality of systematic reviews (Parums, 2021). PRISMA 2020 has been used as a guideline to review past literature systematically in global and local contexts across many subjects (Sabtu & Mohd Matore, 2024; Sakaria *et al*., 2023). In the context of this study, PRISMA 2020 guides conducting a rigorous search of terms related to cognitive diagnostic assessment.

2.1. Formulation of the Research Questions

The formulation of the research questions is guided by the PICo framework (Population or Problem, Interest, and Context), which helps structure systematic reviews by aligning key research components (Lockwood *et al*., 2015). This framework is used to ensure a comprehensive exploration of CDA in language assessment. The Population or Problem (P) refers to CDA studies in language assessment, while the Interest (I) focuses on publication trends, implementation practices, and future research directions. The Context (Co) specifies language assessment settings. Based on this framework, the first research question explores publication trends to map the field's development, the second research question investigates implementation practices of CDA, and the third research question identifies proposed future research directions, offering a systematic understanding of the current state and future possibilities of CDA in language domains.

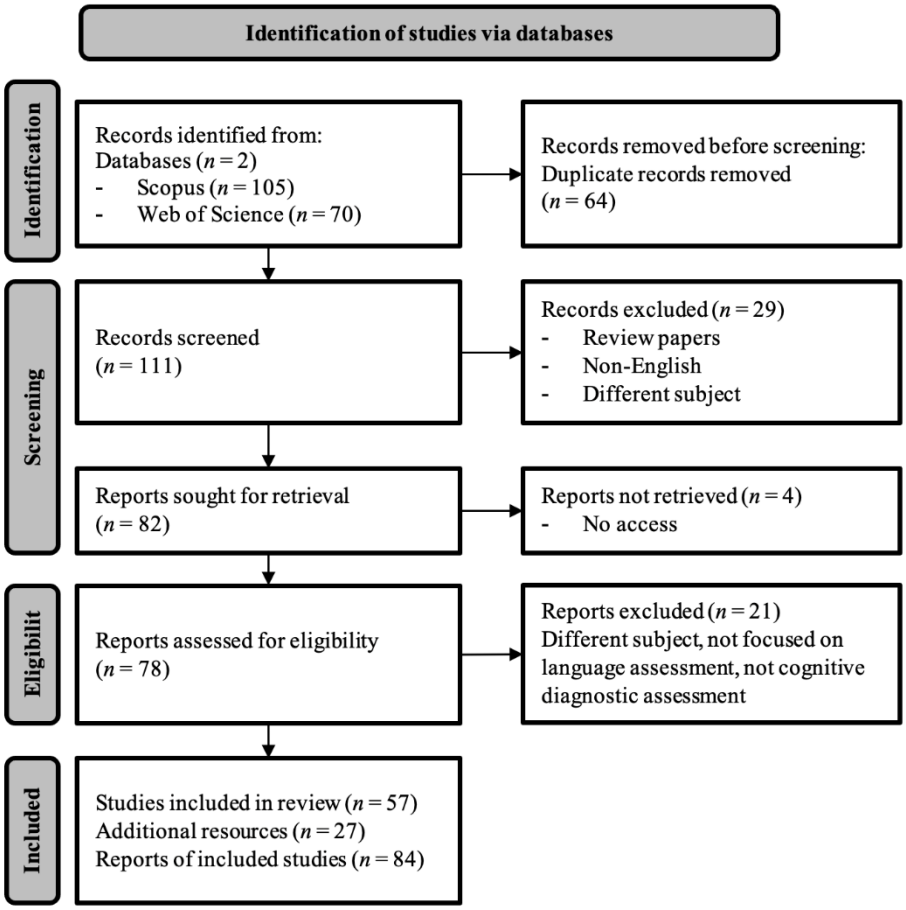
2.2. Database Selection

The articles in this study were sourced from two prominent online databases, specifically Scopus and the Web of Science (WoS). The two databases have been used in past studies as sources to review literature published in a particular field of study (Musid *et al*., 2024; Nadmilail *et al*., 2023). Scopus, a comprehensive database, aggregates content from over 30,000 journals representing a wide range of subject fields, including education, measurement, and language, enabling researchers to identify trends, gaps, and advancements in the field of study (Ishak *et al*., 2025). On the other hand, WoS is a platform providing access to a robust database across numerous academic disciplines (Syed Ismail *et al*., 2024). With its extensive coverage and reputation, WoS serves as a valuable resource for researchers across diverse fields (Mohd Noh *et al*., 2025).

2.3. Review Stages

This review encompassed three principal stages: identification, screening, and eligibility, as presented in Figure 1.

Figure 1. *The PRISMA flow diagram of the study.*



In the initial stage, keywords for the search process were identified, aligning with the research questions and drawing from both previous literature and thesaurus suggestions. The primary keywords, including cognitive diagnostic assessment, cognitive diagnostic model, diagnostic classification model, language and language assessment as shown in Table 1, were selected based on their prominent usage in previous studies. Using the keywords, a total of 105 publications appeared in the Scopus database, while 70 publications appeared in the Web of Science database.

Table 1. *The search string used for the systematic literature review process.*

Database	Keywords
Scopus	TITLE-ABS-KEY (("cognitive diagnostic assessment" OR "cognitive diagnostic model*" OR "diagnostic classification model") AND ("language assessment" OR "language"))
Web of Science	TS = (("cognitive diagnostic assessment" OR "cognitive diagnostic model*" OR "diagnostic classification model") AND ("language assessment" OR "language"))

Then, duplicates were identified, resulting in the removal of 64 duplicate records, leaving a total of 111 articles. Next, screening was implemented to select articles deemed significant for the review (Sohrabi *et al*., 2021), employing various criteria outlined in Table 2. Firstly, articles not in English were filtered out to streamline the review process, eliminating the need for translation. Secondly, only articles presenting empirical data from academic journals were considered, ensuring a focus on studies conducted through well-designed methods. This exclusionary approach left out other publication types such as reports, theses, article reviews, and conceptual papers. Finally, given the systematic literature review's emphasis on language assessment, only articles published within the context of language domains were included. After the screening process, 29 articles were removed, leaving 82 articles for retrieval. However, due to inaccessible access, four articles could not be retrieved, resulting in a total of 78 articles for the next stage.

Table 2. *The filtering criteria.*

Criteria	Included	Excluded
Language	English	Language other than English language
Document types	Journal articles	Conceptual paper, systematic review, book chapter, proceedings, thesis, report
Subject area	Education, social science, humanities	Medicine, Mathematics, Health, Dentistry, Engineering

In the third stage, eligibility criteria were applied. After carefully examining the abstracts, 21 articles were excluded for various reasons, such as focusing on subjects unrelated to language domains, like Mathematics and medicine. Ultimately, 57 articles were selected for review. Additional searches were conducted using backward and forward tracking (Mohamed Shaffril *et al*., 2021). Backward tracking involved reviewing the references cited in the selected publications, while forward tracking involved identifying publications that cited the original article. This process added another 27 articles that met the eligibility criteria, bringing the total number of reviewed publications to 84.

2.4. Quality Appraisal

Quality appraisal is a rigorous evaluative process undertaken to ascertain the methodological quality of studies included in a systematic review, ensuring that only studies adhering to predefined methodological standards are considered (Mohamed Shaffril *et al*., 2024). This process enhances the credibility and accuracy of the review findings by identifying potential methodological weaknesses that may compromise the interpretation of results. The Mixed-Method Appraisal Tool (MMAT), developed by Hong *et al*. (2018), was adopted in this study due to its comprehensive capacity to evaluate various research designs, encompassing quantitative, qualitative, and mixed-methods studies. MMAT serves as a structured framework comprising five key evaluation criteria, each aligned with the specific characteristics of the respective research designs. The MMAT scoring system operates on a scale ranging from 0 to 5, wherein each fulfilled criterion contributes one point to the total score, reflecting the overall methodological robustness of the study. Following the MMAT assessment, all 84 articles

attained a minimum score of three out of the five evaluation criteria and were consequently retained for further analysis.

2.4. Abstraction and Analysis

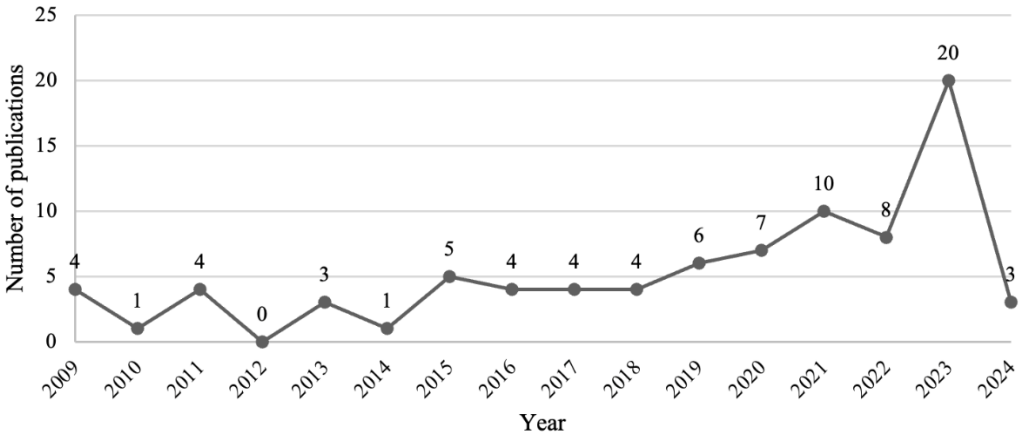
The selected articles underwent a thorough examination and analysis. Initially, their abstracts were reviewed comprehensively. Following this, a deeper exploration of the full articles was conducted to identify possible codes related to the three main themes outlined by the research questions: the research contexts, procedural frameworks and future directions. Each article was meticulously examined, focusing on five criteria: years of publication, geographical locations, research design, focused skills, and research samples. The extracted data were organized and tabulated in Microsoft Excel to generate graphical representations. Then, based on the employed research designs, the articles were classified into either retrofitting studies or true CDA studies. Articles categorized as true CDA studies, which developed new items, were further analyzed by examining their procedural frameworks. Each study phase was compared and contrasted to understand how they addressed the specification of attributes, the design of cognitive models, the development of items, the construction of Q-matrix, the validation and reliability processes, data collection, and data analysis. The data analysis was scrutinized to determine how model fit, item parameters, attribute prevalence, and mastery classes were reported in the selected studies. Finally, future directions were identified through a detailed analysis of the discussion and conclusion sections of the selected publications, where each suggestion or recommendation was further categorized into sub-themes and major themes.

3. FINDINGS

3.1. Findings Related to First Research Question

In this section, findings are presented based on four criteria: years of publications, geographical locations, focused skills and research samples. The four criteria were chosen to provide a comprehensive and multi-faceted analysis of the scope, geographical distribution, specific areas of focus and participant demographics within the field of CDA in a language assessment setting. The growth of research interest in CDA within language assessment can be observed through the increasing number of publications over time. Figure 2 presents the number of publications on cognitive diagnostic assessment in language testing according to the year of publication from 2009 to 2024.

Figure 2. *Publication distribution based on year of publication.*

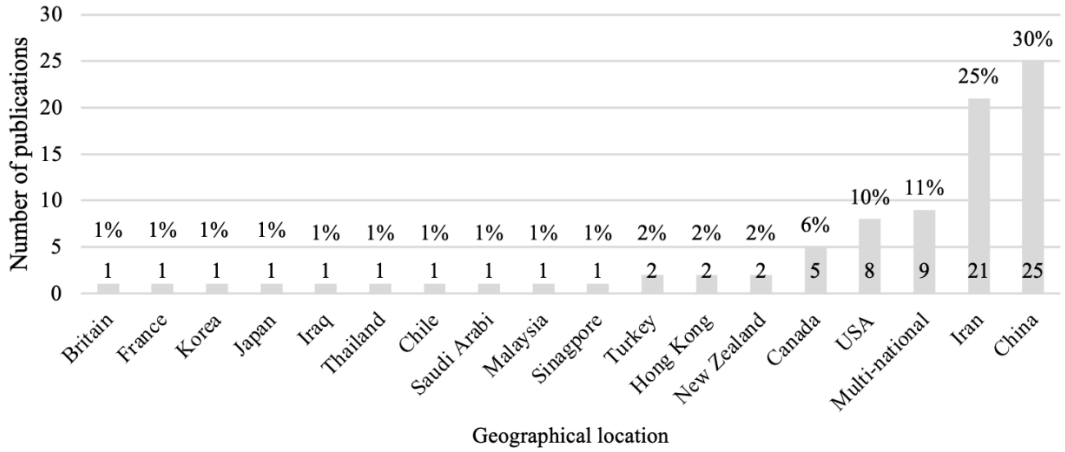


The initial years, 2009 and 2011, both saw a modest number of 4 publications each, accounting for 5% of the total publications per year. This early period shows an initial interest that experienced a slight dip in 2010 with only 1 publication (1%), followed by a return to 4 publications (5%) in 2011. In 2012, there were no publications. However, the following year, 2013, showed a moderate recovery with 3 publications (4%). The number of publications remained relatively low in 2014 with just 1 publication (1%). A noticeable increase occurred in

2015 with 5 publications (6%), marking the beginning of a more consistent period of research interest in the field. The years 2016, 2017 and 2018 recorded 4 publications each (5%), maintaining steady interest. Then, 2019 saw a significant uptick to 6 publications (7%), and the trend continued into 2020 with 7 publications (8%), suggesting sustained interest despite slight fluctuations. The most significant increases occurred in the years 2021 and 2023. In 2021, the number of publications surged to 10 (12%) and was followed by 8 publications in 2022 (10%), showing a slight decline but still maintaining a high level of interest. The year 2023 saw the highest number of publications recorded in the entire period with 20 (24%), indicating a significant and unprecedented spike in research output. As of 2024, there are already 3 publications (4%), although this number may increase as the year progresses. The data reflects a clear upward trend in the number of publications, particularly from 2019 onwards, which suggests a growing recognition of the importance and relevance of the research area.

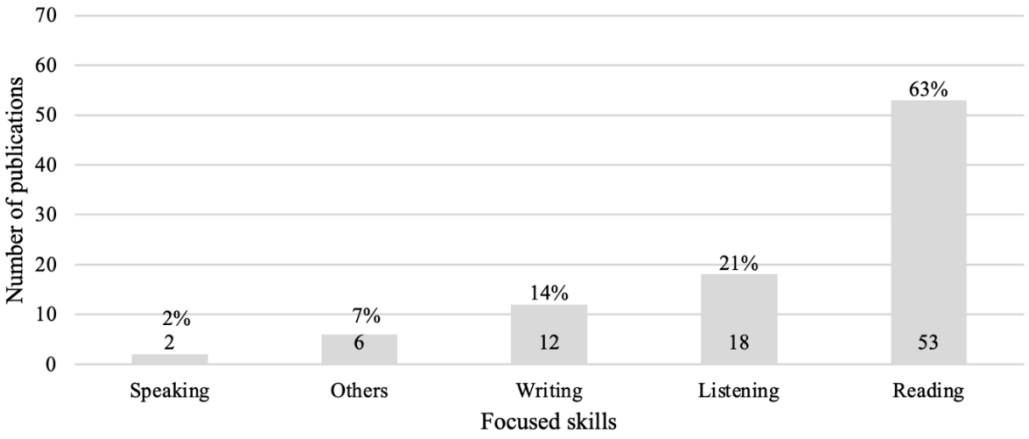
CDA research in language assessment appears across multiple countries and shows differing levels of contribution by region. Figure 3 outlines the publication frequency counts of geographical locations. This distribution reveals a significant variation in research output across different regions. Notably, China and Iran have a higher count of publications, with 25 publications (30%) and 21 publications (25%) respectively. Studies at the international level involving multiple countries follow closely with nine publications (11%). The United States comes next by a substantial margin with eight publications (10%). Countries such as Canada, with five publications (6%), and Turkey, Hong Kong, and New Zealand, each with two publications, demonstrate a moderate level of engagement in this research field. On the lower end, many countries, including Britain, France, Korea, Japan, Iraq, Thailand, Chile, Saudi Arabia, Malaysia, and Singapore have each produced one publication.

Figure 3. *Publication distribution based on geographical locations.*



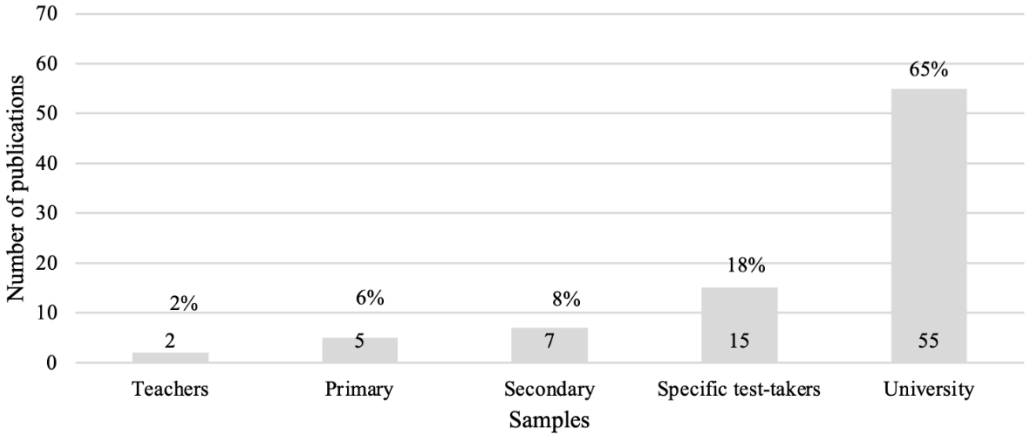
The distribution of CDA studies across different language skills reveals a clear preference in research focus. Figure 4 illustrates the number of publications categorized according to the language skills targeted in CDA studies. While the majority of studies focused on a single skill, three studies investigated three language skills, and another two studies examined two skills. Reading emerged as the most frequently assessed skill, accounting for 53 studies (63%), followed by listening with 18 studies (21%) and writing with 12 studies (14%). A smaller number of studies addressed other skills such as translation and grammar (6 studies, 7%), while speaking was the least represented, with only 2 studies (2%). This distribution highlights a notable research imbalance, with a strong emphasis on receptive skills while productive skills such as speaking remain significantly underrepresented.

Figure 4. *Publication distribution based on focused skills.*



The types of research samples used in CDA studies vary across educational levels and testing populations. Figure 5 depicts the distribution of CDA publications based on the types of research samples used for data collection. The majority of studies, accounting for 55 publications (65%), utilize university students as their samples. These university students are further categorized into groups such as undergraduate students (He *et al*., 2021; Toprak & Cakir, 2021), master's students (Geramipour *et al*., 2021; Tonekaboni *et al*., 2021), and doctoral students (Javidanmehr *et al*., 2019; Shahmirzadi & Marashi, 2023). Additionally, 15 studies (18%) focused on the test-takers of specific tests, including those taking TOEFL (Sawaki *et al*., 2009; Yi, 2016), IELTS (Panahi & Mohebbi, 2022), PIRLS (Thi & Loye, 2019), MELAB (Li *et al*., 2015), and PISA (Chen & Chen, 2016). Another seven studies (8%) used high school students as their sample, while five studies (6%) involved primary school students. Furthermore, two studies (3%) employed English language teachers as their sample.

Figure 5. *Publication distribution based on sample types.*



3.2. Findings Related to Second Research Question

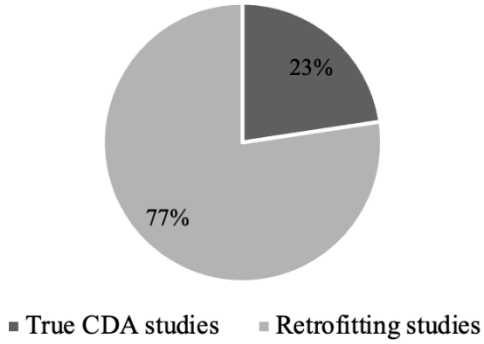
The second research question aims to explore the methodological approaches of CDA in language assessment by examining study designs and statistical analyses.

3.2.1. *Study designs*

CDA studies can be categorized into two main study designs: retrofitting studies and true CDA studies. Retrofitting studies involve the application of cognitive diagnostic models to data derived from pre-existing assessments, while true CDA studies are developed specifically with CDA principles in mind, encompassing item construction, attribute specification, and model alignment from the outset. Each design offers distinct contributions to advancing the precision and effectiveness of language assessment through diagnostic insights. The review found that retrofitting is the dominant approach in cognitive diagnostic assessment studies, with 77% of

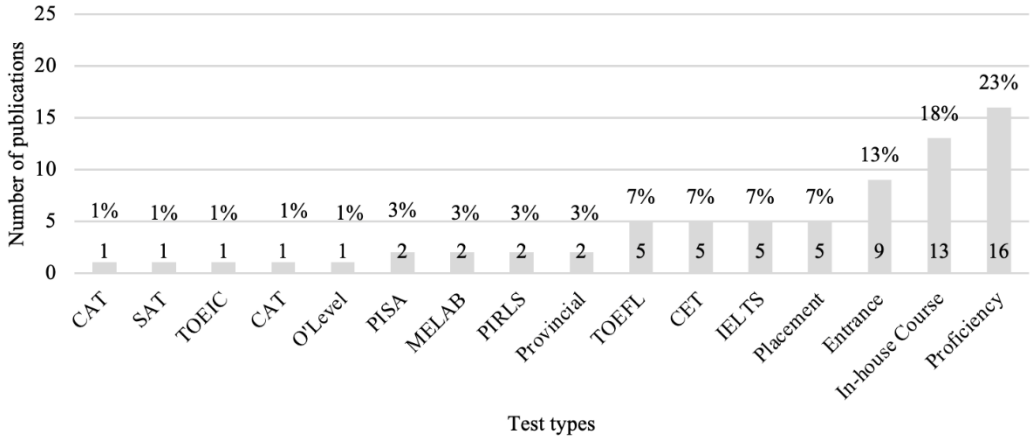
the 84 studies employing this design, while only 23% adopted a true CDA study design, as illustrated in Figure 6.

Figure 6. *Publication distribution based on study designs.*



Retrofitting is an approach to diagnosing and analyzing data from existing testing systems to produce candidates' diagnostic profiles (Li *et al*., 2021). It involves the process of reanalyzing data from existing assessments to extract detailed information about students' specific cognitive strengths and weaknesses. This is achieved by mapping the test items to a cognitive model that specifies the underlying skills or attributes measured by each item. By doing so, researchers can gain insights into individual learners' knowledge states and tailor instruction to address specific areas of need (Sessoms & Henson, 2018). This process may involve improving the test structure, modifying test items, or adjusting diagnostic models. Retrofitting studies in CDA have used various existing language assessments to capture the wide range of test purposes and contexts. Figure 7 presents the types of tests used in previous retrofitting studies.

Figure 7. *Distribution of language test types in retrofitting studies.*



The top three types of language tests are proficiency tests (16 studies, 23%), tests for in-house courses (13 studies, 18%), and entrance examinations for university programs (10 studies, 13%). Although these three types of tests are developed by individual educational institutions such as universities, they are designed for different purposes. Proficiency tests are designed to assess overall language ability. Entrance examinations are used to select students for enrolment in academic programs. Meanwhile, in-house language course tests aim to assess students' understanding and progress within a specific course. Additionally, a significant number of international language tests are used, including the Test of English as a Foreign Language (TOEFL), the International English Language Testing System (IELTS), and the Chinese English Test (CET), each with five studies (7%).

Unlike retrofitting study design, true cognitive diagnostic assessment studies develop new testing items based on CDA approaches from the ground (Ravand & Baghaei, 2019). These types of studies undergo the whole process of development and validation of new items. However, the implementation of true studies is likely chosen over studies through the

retrofitting method due to its complexity and comprehensiveness. To understand how CDA items for language assessment were developed, this review examines the procedural framework undertaken in true CDA studies. Nine studies were identified and selected, each employing different procedures as shown in Table 3. Seven of these studies focused on reading skills, while two on writing skills. The studies were compared based on the phases of attribute specification, item development and Q-matrix construction.

All these true CDA studies began with attribute specification to identify the specific cognitive attributes that accurately measure the skill under study. Various methods were used for attribute specification, including literature reviews (Li *et al*., 2021), think-aloud protocols (Askari & Karami, 2024), modifications of existing rating checklists (Shi *et al*., 2024), expert panel reviews (Nallasamy & Khairani, 2022), exam report analysis, and self-assessment surveys (Toprak & Cakir, 2021). A comprehensive attribute specification process was conducted by Toprak and Cakir (2021), resulting in the generation of a cognitive model as a framework for developing assessment items. The specification of attributes or the development of a cognitive model/ framework is a crucial phase in CDA studies as it ensures that the assessment items accurately measure the intended cognitive skills and processes (Ketabi *et al*., 2021).

The number of attributes specified in the studies under review varies, with the highest being nine attributes (Alavi & Ranjbaran, 2018) and the lowest being three attributes (Lee, 2023). Based on the specified attributes, assessment items were developed, with different studies opting for distinct item types. All studies focusing on reading skills chose the multiple-choice format. Four of these studies constructed the items based on reading passages as stimuli. The highest number of passages used was five (Toprak & Cakir, 2021), followed by four passages (Nallasamy & Khairani, 2022; Puente *et al*., 2023) and three passages (Li *et al*., 2021; Ranjbaran & Alavi, 2017). The largest number of items was developed by Li *et al*. (2023), who created 160 items to build an item bank for computer-adapted testing. Items from other studies ranged from 12 (Puente *et al*., 2023) to 27 (Toprak & Cakir, 2021). For studies on writing skills, Lee (2023) used one writing prompt, while Shi *et al*. (2024) used three writing prompts.

The next procedure is the construction of a Q-matrix to map the relationship between items and attributes. Q-matrix is an important element of CDA studies as it outlines the attributes required to answer each test item accurately. Q-matrix was first introduced by Tatsuoka (1983, 1990) and is defined as an incidence matrix that maps the relationship between test items and the attributes intended to be measured. Generally, the matrix is a matrix between y (number of items) by x (number of attributes), reflecting which attributes are measured by each item (Nájera *et al*., 2020). This matrix summarizes the relationship between items and attributes in a binary manner, with 1 indicating that an item measures the attribute and 0 indicating that it does not (Shan & Wang, 2020). Therefore, each item will take a value of 1 or 0, indicating which attribute is relevant to correctly answering the item. By determining this relationship, the Q-matrix allows for a detailed analysis of students' mastery levels of different skills based on their responses to individual items.

All nine studies have developed the Q-matrix to map the tentative relationship between the cognitive attributes and the assessment items. The matrix was then validated either qualitatively through expert review (Li *et al*., 2021) or quantitatively through empirical data analysis (Ranjbaran & Alavi, 2017). The validation process in CDA studies is executed using a number of methods to ensure the accuracy of the developed items and the interim Q-matrix. Pilot testing was conducted to determine how effectively the items measure the intended constructs. Shi *et al*. (2024) validated their scoring descriptors by piloting the scoring procedures with six raters. Nallasamy and Khairani (2022) conducted a qualitative pilot study of their reading items, through a think-aloud protocol to confirm the attributes and validate the Q-matrix. Meanwhile, Li *et al*. (2021) and Toprak and Cakir (2021) conducted two pilot studies with different samples, followed by item refinement procedures after each pilot. Additionally, content validity was assessed through expert reviews (Li *et al*., 2021; Toprak & Cakir, 2021). Furthermore, Li *et al*. (2021) also reported on the internal, construct, and external validity of the developed items.

Table 3. *Procedural framework of selected true CDA studies.*

Authors/ Skills	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5	Phase 6	Phase 7
Askari and Karami (2024) Reading	Attribute specification (7 attributes)	Items & Q-matrix development	Data collection (301 university students)	A single CDM analysis	Report mastery classes	-	-
Shi, *et al*. (2024) Writing	Attribute specification (5 attributes)	Items & Q-matrix development	Pilot test	Data collection (1166 first-university students & 6 raters)	Multiple CDMs analysis	Report mastery classes	-
Puente, *et al*. (2023) Reading	Item development	Data collection (579 university students)	Multiple CDMs analysis	Report mastery classes	-	-	-
Lee (2023) Writing	Attribute specification (3 attributes) and Q-matrix design	Item development	Data collection (315 university students & 2 markers)	A single CDM analysis	Report mastery classes	-	-
Li, *et al*. (2023) Reading	Item development and calibration (6 attributes)	Data collection (28,485 school students)	Multiple CDMs analysis	Report mastery classes	-	-	-
Nallasamy and Khairani (2022) Reading	Attribute specification (4 attributes)	Item development	Pilot test	Data collection (900 school students)	A single CDM analysis	Report mastery classes	-
Li, *et al*. (2021) Reading	Attribute specification (6 attributes)	Items & Q-matrix development	Expert validation of items and Q-matrix &	2 pilot tests	Data collection (21,466 school students)	Multiple CDMs analysis	Report mastery classes
Toprak & Cakir (2021) Reading	Cognitive model development (5 attributes)	Items & Q-matrix development	Expert validation of items & Q-matrix	2 pilot tests	Data collection (1,058 university students	A single CDM analysis	Report mastery classes
Alavi and Ranjbaran (2017) Reading	Attribute specification (9 attributes) & item development	Data collection (1,986 university students)	Q-matrix construction & validation	A single CDM analysis	Report mastery classes	-	-

3.2.2. *Statistical analysis*

Regarding statistical analysis, the majority of the studies utilized a single cognitive diagnostic model (CDM) such as the fusion model (Askari & Karami, 2024; Ranjbaran & Alavi, 2017), the log-linear cognitive diagnostic model (Lee, 2023; Toprak & Cakir, 2021), and the G-DINA model (Nallasamy & Khairani, 2022). One study opted to compare CDM and IRT-based analysis (Puente *et al*., 2023). Other studies compared multiple CDMs to determine the best fit for the data (Li *et al*., 2021, 2023; Shi *et al*., 2024). Additionally, some studies integrated CDM analysis with other types of analysis, such as multi-faceted Rasch measurement (Shi *et al*., 2024) and analysis of variance (Askari & Karami, 2024). Based on the studies under review, there are several significant analyses reported, including model fit, item parameters, attribute prevalence and mastery class profiles. The majority of the studies have reported the model fit which was reported based on the average difference between the estimated data by the model and the observed data collected from the study samples. The analysis of model fit depends on the types of CDM chosen. Assessing model fit is crucial to ensure that the model provides useful and accurate findings.

Like other statistical models, a CDM has no value if it does not fit the available data. There are two main ways to assess the fit of a CDM: relative fit indices and absolute fit indices (Zhang & Wang, 2020). The reviewed studies have measured absolute fit through various indices including Square Root of Mean Square Residuals (SRMSR) (Askari & Karami, 2024), fit chi-square (Puente *et al*., 2023), bivariate fit (Toprak & Cakir, 2021), FUSIONStats and item mastery statistics (IMStats) (Ranjbaran & Alavi, 2017), root mean square error of approximation (RMSEA) (Shi *et al*., 2024), bivariate Pearson Chi-Square and the bivariate log-likelihood Chi-Square (Lee, 2023). Whereas, other studies have analysed the relative fit using the three common indices of relative fit indices, -2 log-likelihood (-2LL), Akaike's Information Criterion (AIC) (Akaike, 1974), and Bayesian Information Criterion (BIC) (Schwarz, 1978) (Li *et al*., 2021; Nallasamy & Khairani, 2022; Shi *et al*., 2024) as these studies compared the performance of several CDMs before deciding which model best suit the data.

The studies under review have reported different indices of item parameters, including difficulty index, discrimination index, completeness index, guessing and slipping parameter. The difficulty index has been reported by Lee (2023), Li *et al*. (2023), Puente *et al*. (2023) and Ranjbaran and Alavi (2017) to identify which items are more or less accessible to respondents with varying levels of skill mastery. Next, discrimination was documented by Lee (2023), Li *et al*. (2023), Ranjbaran and Alavi (2017), Toprak and Cakir (2021) and Shi *et al*. (2024) to show clear distinctions in response patterns between those who have mastered the skill and those who have not. Subsequently, the completeness index was only reported by Ranjbaran and Alavi (2017) to ensure that no critical skill is left unmeasured. The guessing parameter accounts for the probability that a respondent with no mastery of the relevant skills can correctly answer an item by chance was documented by Li *et al*. (2023) and Nallasamy and Khairani (2022). Slipping parameter which represents the probability that a respondent with full mastery of the relevant skills incorrectly answers an item due to carelessness, misunderstanding, or other factors was employed only by Nallasamy and Khairani (2022).

Attribute prevalence indicates test-takers mastery probability of each attribute (Shi *et al*., 2024). The higher the value of an attribute prevalence is, the easier attributes are (Xie, 2016). This concept is crucial for understanding the distribution of skills among the study samples and for interpreting the results of the diagnostic assessment. Attribute prevalence was reported in some studies by comparing the extent to which different attributes were mastered by their samples (Askari & Karami, 2024; Nallasamy & Khairani, 2022; Ranjbaran & Alavi, 2017; Shi *et al*., 2024). An attribute mastery class provides a comprehensive depiction of an individual's proficiency across specific attributes. This profile is constructed through the analysis of response data obtained from diagnostic tests, where each attribute corresponds to a distinct cognitive skill or knowledge component. The number of latent classes depends on the number

of attributes used in the study to measure the targeted skills. Most of the reviewed studies reported the mastery class either by selecting the significant classes or presenting all the possible classes. Shi *et al*. (2024) reported 10 dominant classes, Li *et al*. (2023) presented 15 dominant classes, Li *et al*., (2021) selected 15 dominant classes and Toprak and Cakir (2021) reviewed 14 dominant classes. Whereas, Lee (2023), Nallasamy and Khairani (2022) and Puente *et al*. (2023) disclosed all the 16 mastery classes.

3.3. Findings Related to Third Research Question

The third research question focuses on determining future research directions proposed in the cognitive diagnostic assessment studies of language assessment. The analysis of recommendations derived from previous studies on CDA in language skills can be categorized into seven key domains: items and instrumentation, attributes, population and samples, usefulness and effectiveness of CDA, cognitive diagnostic models, development and refinement of Q-Matrix, and cognitive diagnostic analysis. Additionally, another seven recommendations are classified under the miscellaneous category, encompassing aspects such as reliability and validity, development of authentic cognitive diagnostic assessment items, applicability of CDA across different domains, utilization of diverse data collection methods, employment of various study designs, collection of students' demographic profiles, and comparison of students' reading skills mastery with other relevant variables. Table 4 presents the recommendations extracted from the studies under review.

3.3.1. *Items and instruments*

The design of test items and instruments plays a pivotal role in the effectiveness of CDA, as the quality, type, and distribution of items directly impact the accuracy of attribute classification and diagnostic feedback. Previous studies have made several recommendations to enhance item design, selection, and instrument construction in CDA to ensure that tests measure a comprehensive range of cognitive skills. One recurring recommendation is the need to ensure a balanced distribution of items across different attributes. Studies in CDA should emphasize having an equal number of items representing each attribute to improve the accuracy of diagnostic classification by preventing certain skills from being over- or under-represented (Min *et al*., 2022; Min & He, 2022). Additionally, more items targeting multiple attributes simultaneously are necessary as they could provide richer diagnostic information (Li *et al*., 2023). Increasing the number of multi-attribute items is particularly important in tasks like reading comprehension, where multiple cognitive processes are involved (Dong *et al*., 2021; Ravand & Robitzsch, 2018). To improve the discriminatory power of CDA tests, several studies recommend including items with varying levels of difficulty. The inclusion of difficult items can prevent ceiling effects, where advanced learners achieve full mastery without showing the variation needed for accurate diagnosis (Kim, 2015; Seipel *et al*., 2023).

Furthermore, Jang (2009b) and Zhao *et al*. (2020) suggest that item design should go beyond factual knowledge to assess higher-order cognitive skills such as critical thinking and inferencing. The format of test items also affects the diagnostic power of CDMs. Hemati and Baghaei (2020) highlight the use of different item types, such as multiple-choice questions, gap-filling tasks, and open-ended questions, as they may yield different levels of diagnostic accuracy. Whereas, Jang (2009a) emphasizes the need to investigate how the number and order of questions influence CDA results. Another key recommendation is to conduct CDA studies across multiple test forms to ensure that findings are consistent and generalizable. Min *et al*. (2022) recommend replicating studies across different versions of listening tests, while Sawaki *et al*. (2009) suggest using larger numbers of test forms to improve the robustness of diagnostic inferences. These recommendations emphasize the need for replication and cross-validation in CDA research.

Table 4. *Future research direction recommended by previous studies.*

Themes	Categories
Items and instruments	Balanced item distribution and attribute representation Item difficulty and cognitive complexity Item type and format Test form variety and stability
Attributes	Attribute definition and selection Grain size and attribute granularity Attribute interaction and hierarchical structure Model interpretability and stability Practical implementation of attributes in diagnostic assessment
Population and samples	Sample size and stability Population diversity and representativeness Generalizability across contexts and grade levels
Usefulness and effectiveness of CDA	Impact on pedagogical practices Effectiveness of CDA feedback in guiding remedial learning and instruction Stakeholder engagement and perceptions of diagnostic feedback Validity and accuracy of CDA feedback in teaching and learning contexts
Cognitive diagnostic models	Selection and comparison of CDMs Model improvement and development Diagnostic inference and model assumptions Practical considerations in model selection and implementation
The development and refinement of Q-Matrix	Methods, strategies, and automatic revision for Q-matrix development Objectivity and balance in Q-matrix construction Evaluating Q-matrix quality and its effect on diagnostic accuracy
Cognitive diagnostic analysis	Model fit and statistical indices Diagnostic accuracy and subskill classification Item distribution and quality
Others	Reliability and validity, developing new CD assessment items, the applicability of CDA to other domains, the use of diverse data collection methods, employment of different study designs, collection of the demographic profile of students, comparison of students' reading skills mastery to other variables

3.3.2. *Attributes*

The role of attributes in CDA is critical in providing diagnostic feedback on learners' strengths and weaknesses as they represent fine-grained cognitive skills or subskills. Previous studies have proposed various recommendations to enhance the definition, selection, granularity, interaction, and practical application of attributes. Defining and selecting appropriate attributes is the foundation of CDA, as the accuracy of diagnostic feedback depends largely on the relevance and sufficiency of selected attributes. Several studies have highlighted the need to address limitations in attribute coding procedures and to improve attribute selection processes. Svetina *et al*. (2011) suggested further research on the stability of attribute probability estimates in cases of wrongful coding, as coding errors can compromise the reliability of diagnostic outcomes. Construct underrepresentation, the failure to include all necessary cognitive skills was raised by Doe (2014) as a significant concern, indicating the need for comprehensive attribute coverage. Other studies recommended the exploration of additional attributes (Alavi & Ranjbaran, 2018; Ranjbaran & Alavi, 2017), particularly through expert judgment or data-driven methods (Li *et al*., 2021). One of the most debated issues in CDA is the grain size of attributes, and whether attributes should be defined at a coarse (broad) or fine (detailed) level. The grain size directly affects both the diagnostic precision and efficiency of feedback. Boori *et al*. (2024) emphasized that finer-grained attributes could provide more detailed diagnostic information but might increase the complexity of models and reduce classification accuracy. In contrast, merging smaller attributes into composite attributes can simplify the model without sacrificing interpretability (Tonekaboni *et al*., 2021). Several studies (Li & Suen, 2013; Ravand,

2016) recommended balancing theoretical, technical, and practical considerations when determining the appropriate grain size.

While many studies have focused on isolated attributes, increasing attention has been given to the interactions between attributes and their hierarchical organization. Cognitive skills are often interdependent, with certain subskills serving as prerequisites for others. For example, Yi (2016) proposed exploring how reading subskills differ in their contribution and difficulty levels, while Chen *et al*. (2023) suggested investigating how mastering certain subskills facilitates the use of others. Hierarchical models, such as those proposed by Min and He (2022) for listening subskills, offer a more realistic representation of language abilities by capturing the progressive nature of skill acquisition. Additionally, expert-based insights into attribute interaction could further refine diagnostic models (Yumsek, 2023). A recurring challenge in CDA is ensuring that diagnostic results are interpretable and stable across different contexts. Interpretability refers to the extent to which diagnostic profiles provide meaningful insights into learners' cognitive abilities. Lee and Sawaki (2009) emphasized the importance of understanding less common skill mastery profiles, particularly when learners exhibit unusual combinations of skills. Meanwhile, Svetina *et al*. (2011) pointed out the need for simulation-based studies to assess the stability of attribute probability estimates under conditions of coding error. Both studies highlight the importance of improving the transparency and robustness of diagnostic outcomes, particularly when CDMs are applied in real-world educational settings. Translating theoretical models into practical applications is essential for the widespread adoption of cognitive diagnostic assessment. Mohammed *et al*. (2023) advocated for retrofitting existing assessments to extract cognitive diagnostic information before designing entirely new diagnostic tests, a cost-effective approach that could accelerate the integration of CDMs into large-scale assessments. Additionally, Lee (2023) explored how self-regulatory strategies could be incorporated into diagnostic models, suggesting that attributes related to metacognitive processes might provide richer insights into learners' performance.

3.3.3. *Population and samples*

Previous studies have provided various recommendations regarding population coverage and sample selection to improve the robustness of diagnostic outcomes. Several studies emphasize that larger sample sizes are essential to ensure consistent and stable diagnostic inferences in CDA. Smaller sample sizes often yield unstable attribute probability estimates, making the diagnostic results less reliable. For example, Jang (2009a) and Mohammed *et al*. (2023) both contended for increasing sample sizes to enhance the accuracy of mastery profiles. Shahmirzadi (2023) emphasized that data analysis could benefit from a bigger sample size to achieve more reliable interpretations of cognitive processes. Diverse population samples contribute to the external validity of CDA findings, making the results more generalizable across different learner populations. Studies like Shahmirzadi (2023) and Wang *et al*. (2024) recommended collecting data from multiple institutions and diverse educational settings to allow for broader comparisons. Similarly, Meng *et al*. (2023) suggested including learners with different proficiency levels to investigate how cognitive attributes vary across populations. Extending the applicability of CDA models across different educational contexts and grade levels is essential for broader implementation. Poolsawad *et al*. (2015) highlighted the need to explore how diagnostic models perform across different grade levels, while Lee and Sawaki (2009) highlighted the importance of testing the stability of content across multiple test forms. These recommendations suggest that diagnostic models should be adaptive and flexible to cater to a wide range of learner populations.

3.3.4. *Usefulness and effectiveness of CDA*

The usefulness and effectiveness of CDA in language assessment is a significant area of inquiry, as CDA aims to provide detailed diagnostic feedback that helps learners, teachers, and other stakeholders improve language learning outcomes. Several studies emphasized investigating

the transformative potential of CDA in shaping classroom practices. Kim (2015) and Toprak and Cakir (2021) recommended examining how CDA feedback can provide teachers with specific insights into students' cognitive strengths and weaknesses. Sun and Hwang (2023) highlighted the need to assess the long-term effectiveness of CDA, as the immediate benefits of diagnostic feedback may differ from its long-term impact on language development. However, the successful integration of CDA into pedagogical practices requires ongoing teacher training and support systems to help educators interpret and apply diagnostic feedback effectively (Shahsavar, 2019). CDA has been widely recognized for its potential to guide remedial teaching and individualized instruction. Previous studies suggested proving how diagnostic feedback can help teachers identify specific skill gaps and design targeted interventions (Li & Wang, 2017; Min & He, 2022). It was also advocated that deeper student engagement with diagnostic feedback could enhance learning outcomes. However, more research is needed to explore how CDA feedback can be tailored to different learner profiles and how learners' responses to feedback affect their long-term skill development (Jang *et al*., 2015). One significant area of research involves the engagement of multiple stakeholders, including students, teachers, and parents, in the diagnostic feedback process. Aryadoust (2018) and Boori *et al*., (2024) emphasized that diagnostic information could help stakeholders set more realistic educational goals. However, Li *et al*. (2021) and Shi *et al*. (2024) pointed out the importance of investigating how learners perceive and interpret diagnostic feedback, as negative perceptions could hinder its effectiveness. The active involvement of stakeholders in co-creating diagnostic reports could improve both the acceptability and utility of diagnostic information.

3.3.5. *Selection and improvement of cognitive diagnostic models*

The models used in cognitive diagnostics play a central role in diagnosing learners' cognitive strengths and weaknesses by mapping their performance to predefined attributes. Different CDMs have been proposed to improve the accuracy, flexibility, and applicability of CDA in language assessment. Previous studies have provided several recommendations to enhance the selection, development, and implementation of CDMs. Selecting the appropriate CDM is one of the most critical decisions in CDA research, as different models vary in their assumptions, diagnostic capabilities, and classification accuracy. Recommendations made include conducting comparative analyses to evaluate the performance of different models (Dong *et al*., 2021; Ravand & Robitzsch, 2018), developing models that incorporate rater variations as a facet, especially for polytomous attributes (Xie, 2016) and establishing simpler calibration techniques that allow CDMs to be used in classroom-based assessments without requiring large-scale participants (Aryadoust, 2011).

3.3.6. *The development of Q-Matrix*

The accuracy of cognitive diagnostic models largely depends on how well the Q-matrix represents the relationship between test items and the target attributes. However, Q-matrix construction poses several methodological challenges, including subjectivity in item-attribute mapping, lack of standardization, and misclassification risks. Previous studies have offered several recommendations to improve Q-matrix construction, with a focus on enhancing its precision, objectivity, and diagnostic accuracy. The matrix should be constructed alongside the development of diagnostic tests rather than retrofitting existing tests, as this approach ensures that item-attribute relationships are more accurately specified from the outset (Hemati & Baghaei, 2020). The accuracy of item-attribute mappings can be further enhanced by incorporating multiple perspectives through the involvement of think-aloud protocols and expert panels (Toprak-yildiz, 2021). Additionally, automatic revision algorithms could refine Q-matrix structures by identifying misclassified items and adjusting attribute mappings based on empirical evidence (Chen & Chen, 2016). On top of that, subjectivity remains a persistent issue in Q-matrix construction, particularly when item-attribute mappings are based solely on expert judgment. Q-matrix development should integrate empirical data sources alongside

expert opinions to improve objectivity (Shahmirzadi, 2023). Balanced Q-matrix designs are essential to ensure that all cognitive attributes are represented equally, preventing certain skills from being overemphasized while others are neglected (Li *et al*., 2021).

3.3.7. *Analysis*

Psychometric considerations include issues related to model fit, subskill classification, item quality, and diagnostic accuracy. One of the key psychometric challenges in CDA is ensuring that diagnostic models accurately fit the data. Model fit reflects how well the estimated model represents the observed item responses. Addressing issues related to item distribution, model convergence, and estimation techniques are essential to improve model fit (Toprak-yildiz, 2021). Exploring alternative statistical fit indices and establishing clearer criteria for fit measures in CDMs were also recommended (Du & Ma, 2021; Yi, 2017). Sensitivity analyses could provide more robust subskill classifications by examining how different model specifications affect diagnostic outcomes (Min *et al*., 2022). Additionally, diagnosing weaknesses at the option level, rather than solely at the item level, could provide more granular diagnostic feedback (Meng *et al*., 2023). Research on flat mastery profiles, where learners are assigned the same mastery status across multiple attributes, is also recommended as this issue could undermine the discriminatory power of diagnostic models (Lee, 2023).

3.3.8. *Others*

Previous research has also proposed several recommendations to enhance the validation and reliability of CDA in language diagnosis. Emphasizing the significance of test construction and validation plays a pivotal role in strengthening the overall assessment framework (Dong *et al*., 2021). Furthermore, the incorporation of multiple sources of validity evidence is essential to improve the effectiveness, efficiency, and practicality of the diagnostic approach (Panahi & Mohebbi, 2022). The continuous validation of diagnostic feedback (Fan & Yan, 2020) and the ongoing refinement of diagnostic methods (Poolsawad *et al*., 2015) are equally vital in enhancing the accuracy and utility of CDA. Other notable recommendations included improving the reliability and validity evidence (Fan & Yan, 2020; Panahi & Mohebbi, 2022), developing new diagnostic assessment items instead of retrofitting existing tests to better align with diagnostic purposes (Hemati & Baghaei, 2020; Tonekaboni *et al*., 2021), and exploring the applicability of CDA to other domains, such as thinking skills and higher-order cognitive skills (Meng & Fu, 2023; Zhao *et al*., 2020). Additionally, researchers advocated for the use of diverse study designs and data collection methods, such as think-aloud verbal protocols and cognitive surveys to improve item-attribute mapping accuracy (Chen *et al*., 2023; Tabatabaee-Yazdi, 2020). The collection of students' demographic profiles was also recommended to facilitate further comparisons between reading skill mastery and other learner-related variables (Hemati & Baghaei, 2020; Ravand, 2016).

4. DISCUSSION

This review addresses the three research questions by identifying the publication trends in CDA studies within the context of language assessment, examining their implementation through retrofitting and true study approaches and determining future research directions recommended by previous research. The review indicates that there are more retrofitting studies than true studies in CDA, likely due to the cost-effectiveness and efficiency of using existing assessments to enhance diagnostic feedback (Liu *et al*., 2018). Retrofitting is particularly beneficial when time constraints lead to not-reached items, as it allows for the development of missing data models that improve diagnostic feedback and item parameter accuracy (Liang *et al*., 2022). Additionally, retrofitting enhances traditional assessments by linking psychometrics with cognitive science, thus providing more detailed information about examinees and the quality of assessment tools (von Davier & Lee, 2019). However, retrofitting may not always yield optimal results if existing tests or models do not fully align with diagnostic needs (Boori *et al*., 2023). While retrofitting is a quick and cost-effective way to obtain diagnostic tests, it requires careful

evaluation to ensure reliability and validity (Akbay, 2021). Challenges of retrofitting include difficulties in developing cognitive models, finding enough items to assess attributes, and achieving strong alignment between cognitive models and test data, which can result in inadequate data for psychometric analysis (Gierl & Cui, 2008).

The review has found that CDA and CDMs are more frequently applied to receptive skills, especially reading, rather than productive skills. This focus is likely because receptive skills are more passive and easier to assess cognitively. Receptive skills involve understanding and interpreting language without active production, making them ideal for measuring specific knowledge structures and cognitive processing abilities (Ha, 2021). Assessing receptive skills also provides valuable insights into students' cognitive strengths and weaknesses, aiding in the development of tailored teaching strategies (Aryadoust, 2018). In contrast, productive skills assessments, such as speaking, are more complex and require raters (Mohd Noh & Mohd Matore, 2022). Rater-mediated assessments introduce variability and challenges in maintaining consistent measurement (Mohd Noh & Mohd Matore, 2020). Additionally, receptive and productive skills are interconnected; for example, developing reading skills can improve writing abilities. Therefore, focusing on receptive skills in cognitive diagnostic assessments allows for a comprehensive evaluation of students' language comprehension and cognitive processes. Apart from that, the review suggests that attribute specification in cognitive models can be further detailed. A cognitive model represents the mental processes and structures involved in completing a task or test question and serves as a framework for the assessment process. It helps test developers organize assessment items to accurately measure the intended skills (Toprak-Yildiz, 2022). This model includes detailed definitions of the constructs being evaluated, outlining cognitive processes, skills, and their relationships (Zhang *et al*. 2024). By defining these components, the cognitive model provides a solid foundation for test development, score interpretation, and validation (Fan *et al*., 2021). The model specifies what is to be measured and details the cognitive processes and components an individual goes through while answering test questions (Leighton & Gierl, 2007a; Ranjbaran & Alavi, 2017).

The review also indicates that CDA studies can employ either a single CDM or compare multiple models before selecting the best one. Studies using a single model typically evaluate the overall fit of the model to the data using absolute fit indices. Rupp *et al*. (2010) categorize absolute fit into a global model fit and item-level fit. Global model fit assesses how well the entire statistical model fits the observed data, while item-level fit examines the fit of the CDM to the response data for each test item. Absolute fit indices include SRMSR, fit chi-square, bivariate fit, FUSIONStats, item mastery statistics (IMStats), RMSEA, bivariate Pearson Chi-Square, and bivariate log-likelihood Chi-Square. In contrast, studies comparing different models report both absolute and relative fit. Relative fit indices evaluate whether one model fits the data better than another and include -2LL, AIC, and BIC. On top of that, CDMs can produce various parameters for assessment items by leveraging the data's underlying structure and the relationships between observed responses and latent attributes. Common item parameters in CDA studies include the difficulty index and discrimination index. The difficulty index is based on the proportion of respondents who answer an item correctly, while the discrimination index assesses how strongly an item's response pattern correlates with the overall skill levels of respondents. Other parameters include the completeness index, guessing parameter, and slipping parameter. The completeness index ensures that all skills are adequately represented by the test items. The guessing parameter estimates the likelihood of randomly selecting the correct answer based on the item format, and the slipping parameter analyses the response patterns of respondents expected to have mastered the skill.

Future research directions in CDA publications of language assessment have emphasized several critical areas for advancement. One key direction is the refinement of item and instrumentation design by exploring how various item types, text genres, and item-attribute interactions impact cognitive skill elicitation and classification accuracy. The development of

hierarchical models and composite attributes has been proposed to enhance the granularity and interpretability of diagnostic feedback. Additionally, expanding sample diversity and systematically including underrepresented populations are essential to improve the inclusiveness and generalizability of diagnostic outcomes. Collaborative validation efforts involving multiple stakeholders, including parents, teachers, and learners, have been recommended to strengthen the practical utility of CDA. The integration of automatic Q-matrix generation algorithms and data-driven methods could significantly improve the objectivity and transparency of diagnostic models. Moreover, future studies should prioritize the development of sensitivity analyses, alternative fit indices, and option-level diagnostics to further enhance psychometric properties and model fit. Finally, exploring the application of CDA in assessing higher-order cognitive skills and incorporating self-regulatory and task-completion strategies would contribute to a more holistic understanding of language performance and learner-centred assessment practices.

5. CONCLUSION

While the review provides valuable insights into CDA and its applications, several limitations must be considered. First, the findings may be influenced by publication bias, as studies with positive or significant results are more likely to be published, potentially skewing the conclusions. Second, the reliance on existing literature may overlook unpublished studies, dissertations, or grey literature that could offer additional perspectives or findings not captured in the review. Third, the review's scope may not encompass recent advancements or emerging trends in CDA, as the field is continually evolving with new methodologies and technologies. These limitations underscore the need for cautious interpretation of the review's conclusions and highlight the importance of considering a broader range of evidence to inform future research and practice in CDA.

Based on the comprehensive review provided, several future avenues are suggested to enhance the field of CDA. Firstly, there is a need to focus on developing comprehensive cognitive models rather than relying on retrofitting existing assessments. This involves detailing attribute specifications and thoroughly defining the cognitive processes and skills to be measured, which would significantly improve the accuracy and reliability of CDAs. Secondly, balancing the assessment of receptive and productive skills is crucial. While CDA studies on receptive skills should still be the primary focus, future research should also address productive skills such as writing and speaking to provide a more holistic understanding of students' language abilities and cognitive processes of polytomous attributes. More research on receptive skills is still needed to refine the assessment tools and methodologies used to measure these skills accurately. Furthermore, expanding the analysis of item parameters beyond difficulty and discrimination indices is important. Future studies should systematically analyze parameters such as completeness, guessing, and slipping to provide deeper insights into item performance and test effectiveness. Additionally, integrating artificial intelligence and machine learning into CDA can enhance diagnostic accuracy and provide personalized feedback in real-time. Exploring the application of CDA in diverse cultural and linguistic contexts will help understand its adaptability across different educational systems. Investigating the use of CDA in non-traditional learning environments, such as online courses and workplace training programs, presents another valuable research avenue. Finally, effective teacher training programs and strategies for integrating CDA into existing curricula are crucial for successful implementation, ensuring the practical and equitable application of CDA in education.

The review's findings have significant implications for the field of language assessment. Practitioners must critically evaluate the alignment of existing tests with new diagnostic needs to ensure the reliability and validity of the results. It is also important to consider using various item parameters, such as completeness, guessing, and slipping, to achieve comprehensive and accurate assessment outcomes. Additionally, detailed cognitive models are essential in CDA

studies as they provide a robust framework for test development, ensuring precise measurement of intended skills and enhancing the validity of diagnostic results. For stakeholders, the implications of the current review are equally important. Educational policymakers should support the development and implementation of more refined cognitive models and diverse assessment parameters. This will ensure that diagnostic assessments are both accurate and effective in identifying students' strengths and weaknesses. Teachers and educators need to be trained in using these advanced diagnostic tools to tailor their instructional strategies effectively.

In conclusion, the reviewed studies on the development of CDA items in language assessment showcase a rich diversity in their approaches and methodologies. Each study demonstrates a systematic progression through key stages such as attribute specification, item construction, data analysis, and model selection. The comprehensive nature of these studies, encapsulating various facets of CDA development, reflects the evolving landscape and commitment of researchers to advancing the precision and effectiveness of language assessments. As the field continues to evolve across countries and time, the multifaceted approaches observed in these studies contribute to a deeper understanding of the complexities involved in CDA within the context of language assessment.

Acknowledgments

The study was funded by Faculty of Education, Universiti Kebangsaan Malaysia, Ganjaran Penerbitan GP-KO21854 and TAP-KO21854 Universiti Kebangsaan Malaysia. The authors wish to express their sincere gratitude to the Kumpulan Penyelidikan Universiti, Penilaian Pendidikan, Faculty of Education, Universiti Kebangsaan Malaysia, Bangi, for their invaluable support and contributions to this research.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Contribution of Authors

Muhamad Firdaus Mohd Noh: Conceptualization, Methodology, Formal analysis, and Writing-original draft. **Mohd Effendi Ewan Mohd Matore**: Supervision, Analysis, Review and Editing. **Nur Ainil Sulaiman**: Supervision, Analysis, Review, and Editing.

Orcid

Muhamad Firdaus Mohd Noh https://orcid.org/0000-0002-5429-6789
Mohd Effendi Ewan Mohd Matore https://orcid.org/0000-0002-6369-8501
Nur Ainil Sulaiman https://orcid.org/0000-0001-6212-7494

REFERENCES

Akbay, L. (2021). Impact of retrofitting and item ordering on DIF. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 212-225. https://doi.org/10.21031/epod.886920

Alallo, H.M.I., Mohammed, A., Hamid, Z.K., Hassan, A.Y., & Kadhim, Q.K. (2023). Examining attribute relationship using Diagnostic Classification Models: A mini review. *International Journal of Language Testing, 13* (Special Issue), 21-30. https://doi.org/10.22034/IJLT.2022.368468.1212

Alavi, M., & Ranjbaran, F. (2018). Constructing and validating a Q-Matrix for cognitive diagnostic analysis of a reading comprehension test battery. *Journal of English Language Teaching and Learning, 21*(12), 1-15. https://elt.tabrizu.ac.ir/article_17207_d9a9150ba8996ecfe90af3e07b905503.pdf

Aryadoust, V. (2011). Application of the fusion model to while-listening performance tests. *JALT Testing & Evaluation SIG Newsletter, 15*(2), 2–9.

Aryadoust, V. (2018). A cognitive diagnostics assessment study of the listening test of the Singapore–Cambridge General Certificate of Education O-Level: Application of DINA, DINO, G-DINA, HO-DINA, and RRUM. *International Journal of Listening, 35*(1), 29-52. https://doi.org/10.1080/10904018.2018.1500915

Askari, M., & Karami, H. (2024). On the relationship between sensory learning styles and reading subskill profiles: An application of Fusion Model. *Language Related Research, 15*(3), 245–274. https://doi.org/https://doi.org/10.29252/LRR.15.3.10

Boori, A.A., Ghazanfari, M., Ghonsooly, B., & Baghaei, P. (2023). The construction and validation of a Q-matrix for cognitive diagnostic analysis: The case of the reading comprehension section of the IAUEPT. *International Journal of Language Testing, 13*(Special Issue), 31–53. https://doi.org/10.22034/IJLT.2023.383112.1227

Boori, A.A., Ghazanfari, M., Ghonsooly, B., & Baghaei, P. (2024). A cognitive diagnostic modeling analysis of the reading comprehension section of an Iranian high-stakes language proficiency test. *International Journal of Language Testing, 14*(1), 17-33. https://doi.org/10.22034/IJLT.2023.399561.1256

Chen, H., Cai, Y., & de la Torre, J. (2023). Investigating second language (L2) reading subskill associations: A cognitive diagnosis approach. *Language Assessment Quarterly, 20*(2), 166–189. https://doi.org/https://doi.org/10.1080/15434303.2022.2140050

Chen, H., & Chen, J. (2015). Exploring reading comprehension skill relationships through the G-DINA model. *Educational Psychology, 36*(6), 1049-1064. https://doi.org/10.1080/01443410.2015.1076764

Chen, H., & Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the Generalized DINA Model framework. *Language Assessment Quarterly, 13*(3), 218–230. https://doi.org/10.1080/15434303.2016.1210610

Doe, C. (2014). Diagnostic English Language Needs Assessment (DELNA). *Language Testing, 31*(4), 537–543. https://doi.org/10.1177/0265532214538225

Dong, Y., Ma, X., Wang, C., & Gao, X. (2021). An optimal choice of cognitive diagnostic model for second language listening comprehension test. *Frontiers in Psychology, 12*, 1–12. https://doi.org/10.3389/fpsyg.2021.608320

Du, W., & Ma, X. (2021). Probing what's behind the test score: application of multi-CDM to diagnose EFL learners' reading performance. *Reading and Writing, 34*(6), 1441–1466. https://doi.org/10.1007/s11145-021-10124-x

Embretson, S. (1984). A general latent trait model for response process. *Psychometrika, 49*(2), 175–186.

Fan, T., Song, J., & Guan, Z. (2021). Integrating diagnostic assessment into curriculum: a theoretical framework and teaching practices. *Language Testing in Asia, 11*(2), 1–23. https://doi.org/10.1186/s40468-020-00117-y

Fan, T., & Yan, X. (2020). Diagnosing English reading ability in Chinese senior high schools. *Studies in Educational Evaluation, 67*, 1-12. https://doi.org/10.1016/j.stueduc.2020.100931

Fischer, G.H., & Formann, A.K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement, 6*(4), 397–416. https://doi.org/10.1177/014662168200600403

Geramipour, M., Talebzadeh, H., & Mahdi, S. (2021). The optimal Cognitive Diagnostic Model (CDM) for the Grammar Section of MA Entrance Examination of State Universities for EFL candidates. *Language Related Research, 12*(1), 187-218. https://doi.org/10.29252/LRR.12.1.6

Gierl, M.J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: Interdisciplinary Research & Perspective, 6*(4), 263-268.-https://doi.org/10.1080/15366360802497762

Ha, H.T. (2021). Exploring the relationships between various dimensions of receptive vocabulary knowledge and L2 listening and reading comprehension. *Language Testing in Asia, 11*(1), 1–20. https://doi.org/10.1186/s40468-021-00131-8

He, L., Jiang, Z., & Min, S. (2021). Diagnosing writing ability using China's Standards of English Language Ability: Application of cognitive diagnosis models. *Assessing Writing, 50*, 1–14. https://doi.org/10.1016/j.asw.2021.100565

Hemati, S.J., & Baghaei, P. (2020). A cognitive diagnostic modeling analysis of the English reading comprehension section of the Iranian National University Entrance Examination. *International Journal of Language Testing, 10*(1), 11-32. https://www.ijlt.ir/article_114278_022ab6b64d4bc8acba9c561846828270.pdf

Hong, Q.N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2018). *Mixed Methods Appraisal Tool (MMAT) version 2018: User guide*. McGill Department of Family Medicine.

Huang, Q., & Bolt, D.M. (2023). Relative robustness of CDMs and (M)IRT in measuring growth in latent skills. *Educational and Psychological Measurement, 83*(4), 808–830. https://doi.org/10.1177/00131644221117194

Ishak, H., Matore, M.E.E.M., & Nor, H.A.M. (2025). Bibliometric analysis of adaptive behavior research: Advancing Sustainable Development Goals (SDGs) for children with disabilities. *Journal of Lifestyle and SDG's Review, 5*(2), 1-25. https://doi.org/10.47172/2965-730X.SDGsReview.v5.n02.pe03405

Jang, E.E. (2009a). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing, 26*(1), 31–73. https://doi.org/10.1177/0265532208097336

Jang, E.E. (2009b). Demystifying a Q-Matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly, 6*(3), 210-238. https://doi.org/10.1080/15434300903071817

Jang, E.E., Dunlop, M., Park, G., & Boom, E.H. Van Der. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Language Testing, 32*(3), 1-25. https://doi.org/10.1177/0265532215570924

Javidanmehr, Z., Reza, M., & Sarab, A. (2019). Retrofitting non-diagnostic reading comprehension assessment: Application of the G-DINA Model to a high-stakes reading comprehension test. *Language Assessment Quarterly, 16*(3), 1-18. https://doi.org/10.1080/15434303.2019.1654479

Javidanmehr, Z., & Sarab, M.R.A. (2017). Cognitive diagnostic assessment: Issues and considerations. *International Journal of Language Testing, 7*(2), 73-98. https://www.ijlt.ir/article_114441_feee78c9ca19261a6b44843ca71cd075.pdf

Ketabi, S., Alavi, S.M., & Ravand, H. (2021). Diagnostic test construction: Insights from cognitive diagnostic modeling. *International Journal of Language Testing, 11*(1), 22–35. https://www.ijlt.ir/article_128357_68e74c779145d7be54c24f610a549879.pdf

Kim, A.Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2), 227–258. https://doi.org/10.1177/0265532214558457

Lee, K. (2023). Diagnosing Chinese college-level English as a Foreign Language (EFL) learners' integrated writing capability: A Log-linear Cognitive Diagnostic Modeling (LCDM) study. *Assessing Writing, 57*, 1–13. https://doi.org/10.1016/j.asw.2023.100730

Lee, S., Lee, T., Liu, R., Joo, S., Shi, D., Lee, S., & Lee, T. (2025). Research trends and challenges in diagnostic classification models: Insights from dynamic topic modeling. *Measurement: Interdisciplinary Research and Perspectives*, 1-32. https://doi.org/10.1080/15366367.2025.2493998

Lee, Y.-W., & Sawaki, Y. (2009a). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly, 6*(3), 239–263. https://doi.org/10.1080/15434300903079562

Lee, Y.-W., & Sawaki, Y. (2009b). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly, 6*(3), 172-189. https://doi.org/10.1080/15434300902985108

Leighton, J.P., & Gierl, M.J. (2007a). Cognitive Diagnostic Assessment for Education (1st ed.). Cambridge University Press.

Leighton, J.P., & Gierl, M.J. (2007b). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26*(2), 3-16. https://doi.org/10.1111/j.1745-3992.2007.00090.x

Li, H., Hunter, C.V., & Lei, P.-W. (2015). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing, 33*(3), 391-409. https://doi.org/10.1177/0265532215590848

Li, H., & Suen, H.K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment, 18*(1), 1-25. https://doi.org/10.1080/10627197.2013.761522

Li, L., An, Y., Ren, J., & Wei, X. (2021). Research on the cognitive diagnosis of Chinese listening comprehension ability based on the G-DINA Model. *Frontiers in Psychology, 12*, 1–15. https://doi.org/10.3389/fpsyg.2021.714568

Li, S., & Wang, J. (2017). Diagnostic assessment of L2 Chinese learners' reading comprehension ability. In D. Zhang & C. Lin (Eds.), *Chinese as A Second Language Assessment* (pp. 183–202). Springer.

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

Li, Y., Zhen, M., & Liu, J. (2021). Validating a reading assessment within the cognitive diagnostic assessment framework: Q-matrix construction and model comparisons for different primary grades. *Frontiers in Psychology, 12*, 1-13. https://doi.org/10.3389/fpsyg.2021.786612

Liang, L., Lu, J., Zhang, J., & Shi, N. (2022). Modeling not-reached items in cognitive diagnostic assessments. *Frontiers in Psychology, 13*, 1-16. https://doi.org/10.3389/fpsyg.2022.889673

Lim, Y.S., & Bangeranye, C. (2024). The choice between cognitive diagnosis and item response theory: A case study from medical education. *International Journal of Testing, 24*, 356–389. https://doi.org/10.1080/15305058.2024.2369561

Liu, R., Huggins-Manley, A.C., & Bulut, O. (2018). Retrofitting Diagnostic Classification Models to responses from IRT-based assessment forms. *Educational and Psychological Measurement, 78*(3), 357–383. https://doi.org/10.1177/0013164416685599

Lockwood, C., Munn, Z., & Porritt, K. (2015). Qualitative research synthesis: Methodological guidance for systematic reviewers utilizing meta-aggregation. *International Journal of Evidence-Based Healthcare, 13*(3), 179-187. https://doi.org/10.1097/XEB.0000000000000062

Ma, W., Minchen, N., & de la Torre, J. (2020). Choosing between CDM and Unidimensional IRT: The proportional reasoning test case. *Measurement: Interdisciplinary Research and Perspectives, 18*(2), 87–96. https://doi.org/10.1080/15366367.2019.1697122

Mei, H., & Chen, H. (2022). Cognitive diagnosis in language assessment: A thematic review. *RELC Journal, 55*(2), 1–9. https://doi.org/10.1177/00336882221122357

Meng, Y., & Fu, H. (2023). Modeling mediation in the dynamic assessment of listening ability from the cognitive diagnostic perspective. *Modern Language Journal, 107*, 137-160. https://doi.org/10.1111/modl.12820

Meng, Y., Wang, Y., & Zhao, N. (2023). Cognitive diagnostic assessment of EFL learners' listening barriers through incorrect responses. *Frontiers in Psychology, 14*, 1-11. https://doi.org/10.3389/fpsyg.2023.1126106

Min, S., Cai, H., & He, L. (2022). Application of Bi-factor MIRT and Higher-order CDM Models to an in-house EFL listening test for diagnostic purposes. *Language Assessment Quarterly, 19*(2), 189–213. https://doi.org/10.1080/15434303.2021.1980571

Min, S., & He, L. (2022). Developing individualized feedback for listening assessment: Combining standard setting and cognitive diagnostic assessment approaches. *Language Testing, 39*(1), 90–116. https://doi.org/10.1177/0265532221995475

Mirzaei, A., Vincheh, M.H., & Hashemian, M. (2020). Retrofitting the IELTS reading section with a general cognitive diagnostic model in an Iranian EAP context. *Studies in Educational Evaluation, 64*, 1–10. https://doi.org/10.1016/j.stueduc.2019.100817

Mohamed Shaffril, H.A., Abu Samah, A., Samsuddin, S.F., Ahmad, N., Tangang, F., Ahmad Sidique, S.F., Abdul Rahman, H., Burhan, N.A.S., Arif Shah, J., & Amiera Khalid, N. (2024). Diversification of agriculture practices as a response to climate change impacts among farmers in low-income countries: A systematic literature review. *Climate Services, 35*, 1–18. https://doi.org/10.1016/j.cliser.2024.100508

Mohamed Shaffril, H.A., Samsuddin, S.F., & Abu Samah, A. (2021). The ABC of systematic literature review: The basic methodological guidance for beginners. *Quality and Quantity, 55*(4), 1319–1346. https://doi.org/10.1007/s11135-020-01059-6

Mohammed, A., Kareem, A., Dawood, S., Alghazali, T., Khlaif, Q., Sabti, A.A., & Sabit, S.H. (2023). A cognitive diagnostic assessment study of the Reading Comprehension Section of the Preliminary English Test (PET). *International Journal of Language Testing, 13*, 1-20. https://www.ijlt.ir/article_159609_0730ddc075382ca20884ce32b794c85f.pdf

Mohd Noh, M.F., & Mohd Matore, M.E.E. (2020). Rating performance among raters of different experience through multi-facet Rasch measurement (MFRM) model. *Journal of Measurement and Evaluation in Education and Psychology, 11*(2), 147-162. https://doi.org/10.21031/epod.662964

Mohd Noh, M.F., & Mohd Matore, M.E.E. (2022). Rater severity differences in English language as a second language speaking assessment based on rating experience, training experience, and teaching experience through many-faceted Rasch measurement analysis. *Frontiers in Psychology, 13*, 1–13. https://doi.org/10.3389/fpsyg.2022.941084

Mohd Noh, M.F., & Mohd Matore, M.E.E. (2024). Conundrum and considerations in Cognitive Diagnostic Assessment for language proficiency evaluation. *Information Management and Business Review, 16*(2), 63–72. https://doi.org/10.22610/imbr.v16i2(I).3690

Mohd Noh, M.F., Mohd Matore, M.E.E., & Sulaiman, N.A. (2025). A cross-age odyssey of cognitive reading attributes: A scoping review. *Frontiers in Psychology, 16*, 1-19. https://doi.org/10.3389/fpsyg.2025.1605898

Musid, N.A., Matore, M.E.E.M., & Hamid, A.H.A. (2024). A scoping review on mapping the digital leadership constructs for educational settings: what we can learn? International *Journal of Evaluation and Research in Education, 13*(1), 111-121. https://doi.org/10.11591/ijere.v13i1.26760

Nadmilail, A.I., Mohd Matore, M.E.E., Maat, S.M., & Sheridan, L. (2023). Broad vs. narrow traits: A scoping review of measuring personality traits in teacher selection using the situational judgment test. *Frontiers in Psychology, 14*, 1-15. https://doi.org/10.3389/fpsyg.2023.1217321

Nájera, P., Sorrel, M.A., de la Torre, J., & Abad, F.J. (2020). Improving robustness in Q-matrix validation using an iterative and dynamic procedure. *Applied Psychological Measurement, 44*(6), 431–446. https://doi.org/10.1177/0146621620909904

Nallasamy, R., & Khairani, A.Z. (2022). Validating reading comprehension assessment under the GDINA Model. *Malaysian Journal of Social Sciences and Humanities (MJSSH), 7*(11), 1–12. https://doi.org/10.47405/mjssh.v7i11.1877

Panahi, A., & Mohebbi, H. (2022). Cognitive diagnostic assessment of IELTS Listening: Providing feedback from its internal structure. *Language Teaching Research Quarterly, 29*, 147–160. https://doi.org/10.32038/ltrq.2022.29.10

Parums, D.V. (2021). Editorial: Review articles, systematic reviews, meta-analysis, and the updated preferred reporting items for systematic reviews and meta-analyses (PRISMA) 2020 Guidelines. *Medical Science Monitor, 27*, 1-3. https://doi.org/10.12659/MSM.934475

Poolsawad, K., Kanjanawasee, S., & Wudthayagorn, J. (2015). Development of an English communicative competence diagnostic approach. *Procedia - Social and Behavioral Sciences, 191*, 759–763. https://doi.org/10.1016/j.sbspro.2015.04.462

Puente, A., Alvarado, J.M., Gutierrez de Blume, A., & Calderón Maureira, J.F. (2023). An application of cognitive diagnostic models in the evaluation of reading comprehension in Chilean university students. *Suma Psicológica, 30*(1), 40-47. https://doi.org/10.14349/sumapsi.2023.v30.n1.5

Ranjbaran, F., & Alavi, S.M. (2017). Developing a reading comprehension test for cognitive diagnostic assessment: A RUM analysis. *Studies in Educational Evaluation, 55*, 167–179. https://doi.org/10.1016/j.stueduc.2017.10.007

Ravand, H. (2016). Application of a Cognitive Diagnostic Model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment, 34*(8), 782-799. https://doi.org/10.1177/0734282915623053

Ravand, H., & Baghaei, P. (2019). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing, 20*(1), 24-56. https://doi.org/10.1080/15305058.2019.1588278

Ravand, H., & Robitzsch, A. (2018). Cognitive diagnostic model of best choice: A study of reading comprehension. *Educational Psychology, 38*(10), 1255-1277. https://doi.org/10.1080/01443410.2018.1489524

Rupp, A.A., Templin, J., & Henson, R.A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications.* The Guilford Press.

Sabtu, S.H., & Mohd Matore, M.E.E. (2024). Systematic literature review on the implementation of the Six Sigma approach in education. *International Journal of Evaluation and Research in Education, 13*(1), 262-270. https://doi.org/10.11591/ijere.v13i1.26196

Safari, F., & Ahmadi, A. (2023). Developing and evaluating an empirically-based diagnostic checklist for assessing second language integrated writing. *Journal of Second Language Writing, 60*, 1–15. https://doi.org/10.1016/j.jslw.2023.101007

Sakaria, D., Maat, S.M., & Mohd Matore, M.E.E. (2023). Factors influencing mathematics teachers' pedagogical content knowledge (PCK): A systematic review. *Pegem Eğitim ve Öğretim Dergisi, 13*(2), 1–14. https://doi.org/10.47750/pegegog.13.02.01

Sawaki, Y., Kim, H., & Gentile, C. (2009). Q-Matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly, 6*(3), 190-209. https://doi.org/10.1080/15434300902801917

Seipel, B., Kennedy, P.C., Carlson, S.E., Clinton-Lisell, V., & Davison, M.L. (2023). MOCCA-College: Preliminary validity evidence of a cognitive diagnostic reading comprehension assessment. *Journal of Learning Disabilities, 56*(1), 58-71. https://doi.org/10.1177/00222194221121340

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. https://doi.org/10.1080/15366367.2018.1435104

Shahmirzadi, N. (2023). Validation of a language center placement test: Differential item functioning. *International Journal of Language Testing, 13*(1), 1-17. https://doi.org/10.22034/IJLT.2022.336779.1151

Shahmirzadi, N., & Marashi, H. (2023). Cognitive diagnostic assessment of reading comprehension for high- stakes tests: Using GDINA model. *Language Testing in Focus: An International Journal, 8*(8), 1–16. https://doi.org/10.32038/ltf.2023.08.01

Shahsavar, Z. (2019). Diagnosing English learners' writing skills: A cognitive diagnostic modeling study. *Cogent Education, 6*(1), 1-19. https://doi.org/10.1080/2331186X.2019.1608007

Shan, N., & Wang, X. (2020). Cognitive diagnosis modeling incorporating item-level missing data mechanism. *Frontiers in Psychology, 11*, 1-11. https://doi.org/10.3389/fpsyg.2020.564707

Shi, X., Ma, X., Du, W., & Gao, X. (2024). Diagnosing Chinese EFL learners' writing ability using polytomous cognitive diagnostic models. *Language Testing, 41*(1), 109-134. https://doi.org/10.1177/02655322231162840

Sohrabi, C., Franchi, T., Mathew, G., Kerwan, A., Nicola, M., Griffin, M., Agha, M., & Agha, R. (2021). PRISMA 2020 statement: What's new and the importance of reporting guidelines. *International Journal of Surgery, 88*, 39-42. https://doi.org/10.1016/j.ijsu.2021.105918

Sun, H., & Hwang, Y. (2023). Enhancing EFL learners' English reading performance through the diagnostic feedback of CDA. *Korean Journal of English Language and Linguistics, 23*, 535–553. https://doi.org/10.15738/kjell.23..202307.535

Svetina, D., Gorin, J.S., & Kikumi, K. (2011). Defining and comparing the reading comprehension construct: A cognitive-psychometric modeling approach. *International Journal of Testing, 11*(1), 1–23. https://doi.org/10.1080/15305058.2010.518261

Syed Ismail, S.A., Maat, S.M., & Khalid, F. (2024). 35 years of fraction learning: Integrating systematic review and bibliometric analysis on a global scale. *Eurasia Journal of Mathematics, Science and Technology Education, 20*(12), 1-19. https://doi.org/10.29333/ejmste/15657

Tabatabaee-Yazdi, M. (2020). Hierarchical diagnostic classification modeling of reading comprehension. *SAGE Open, 10*(2), 1–13. https://doi.org/10.1177/2158244020931068

Tatsuoka, K.K. (1983). Rule Space: An approach for dealing with misconceptions based on Item Response Theory. *Journal of Educational Measurement, 20*(4), 345–354.

Tatsuoka, K.K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In Fredericksen, Glaser, Lesgold, & Shafto (Eds.), *Diagnostic Monitoring of Skill and Knowledge Acquisition* (pp. 453–488). Routledge.

Thi, D.T.D., & Loye, N. (2019). Cognitive diagnostic analyses of the Progress in International Reading Literacy Study (PIRLS) 2011 results. *Mesure et Évaluation En Éducation, 42*, 127–166. https://doi.org/10.7202/1084131ar

Thompson, W.J., & Clark, A.K. (2024). Improving instructional decision-making using Diagnostic Classification Models. *Educational Measurement: Issues and Practice, 43*(4), 146–156. https://doi.org/https://doi.org/10.1111/emip.12619

Tonekaboni, F.R., Ravand, H., & Rezvani, R. (2021). The construction and validation of a Q-matrix for a high-stakes reading comprehension test : A G-DINA study. *International Journal of Language Testing, 11*(1), 58-87. https://www.ijlt.ir/article_128361_4488710c7301b77e88a7aca333162624.pdf

Toprak-yildiz, T.E. (2021). An international comparison using cognitive diagnostic assessment: Fourth graders' diagnostic profile of reading skills on PIRLS 2016. *Studies in Educational Evaluation, 70*, 1–10. https://doi.org/10.1016/j.stueduc.2021.101057

Toprak-Yildiz, T.E. (2022). Developing a cognitive model of L2 reading comprehension for evaluative purposes in an EMI context. *Studies in Educational Evaluation, 75*, 1-11. https://doi.org/10.1016/j.stueduc.2022.101200

Toprak, T.E., & Cakir, A. (2021). Examining the L2 reading comprehension ability of adult ELLs: Developing a diagnostic test within the cognitive diagnostic assessment framework. *Language Testing, 38*(1), 106-131. https://doi.org/10.1177/0265532220941470

van der Linden, W.J. (2016). *Handbook of Item Response Theory, Volume 1*. CRC Press.

von Davier, M., & Lee, Y.-S. (2019). *Handbook of Diagnostic Classification Models*. Springer.

Wang, C., & Lu, J. (2021). Learning attribute hierarchies from data: Two exploratory approaches. *Journal of Educational and Behavioral Statistics, 46*(1), 58-84. https://doi.org/10.3102/1076998620931094

Wang, X., Zhang, Q., & Gu, L. (2024). A comparative study on cognitive diagnostic assessment of EFL listening in a Chinese College English Context. *Social Science Research Network (SSRN)*, 1–25.

Xie, Q. (2016). Diagnosing university students' academic writing in English: Is cognitive diagnostic modelling the way forward? *Educational Psychology, 37*(1), 1-22. https://doi.org/10.1080/01443410.2016.1202900

Yi, Y. (2016). Probing the relative importance of different attributes in L2 reading and listening comprehension items: An application of cognitive diagnostic models. *Language Testing, 34*(3), 1–9. https://doi.org/10.1177/0265532216646141

Yi, Y. (2017). In search of optimal cognitive diagnostic model(s) for ESL grammar test data – Findings from a model comparison study. *Applied Measurement in Education, 30*(2), 82–101. https://doi.org/10.1080/08957347.2017.1283314

Yumsek, M. (2023). Educational L2 constructs and diagnostic measurement. *Language Testing in Asia, 13*(1), 1–23. https://doi.org/10.1186/s40468-022-00214-0

Zhang, H., Wu, X., & Ju, M. (2024). Developing a cognitive model of solid geometry based on Interpretive Structural Modeling method. *Heliyon, 10*(5), 1-12. https://doi.org/10.1016/j.heliyon.2024.e27063

Zhang, X., & Wang, J. (2020). On the sequential hierarchical cognitive diagnostic model. *Frontiers in Psychology, 11*, 1–13. https://doi.org/10.3389/fpsyg.2020.579018

Zhao, H., Wang, W., & Huang, Y. (2020). A cognitive analysis of an English reading test through the G-DINA model. *Journal of Physics: Conference Series, 1,* 1-7. https://doi.org/10.1088/1742-6596/1629/1/012037