Information Management and Business Review (ISSN 2220-3796)
Vol. 16, No. 2, pp. 63-72, June 2024

Conundrum and Considerations in Cognitive Diagnostic Assessment for Language Proficiency Evaluation

Muhamad Firdaus Mohd Noh, *Mohd Effendi Ewan bin Mohd Matore
Universiti Kebangsaan Malaysia, Malaysia
muhamad.firdausi@gmail.com, *effendi@ukm.edu.my
*Corresponding Author: Mohd Effendi Ewan bin Mohd Matore

Abstract: Since its first appearance in the field of language testing, cognitive diagnostic assessment (CDA) has attracted attention for its ability to extract the intricacies of students' cognitive abilities. However limited research has discussed the issues in the implementation of CDA. Therefore, this article offers an overview of CDA's implementation in language proficiency evaluation. The article also engages in a comprehensive discussion on the conundrum and considerations within CDA, particularly the ongoing debate between distinct classifications of cognitive diagnostic models. It elaborates on the distinctions between the models and their implications for assessment depth and diagnostic insights. Additionally, this article delves into the clash between retrofitting existing items and developing new diagnostic items, highlighting the strategic considerations in each approach. Apart from that, the contentious issue of validating Q-matrices, crucial in CDA, is thoroughly examined, presenting the battle between expert-based and empirical validation methods. The persistent challenges in CDA have profound implications for both theoretical frameworks and practical applications. The theoretical debate not only influences our understanding of cognitive processes but also shapes the conceptualization of diagnostic information extraction. In practical terms, decisions regarding item development, retrofitting strategies, and Q-matrix validation methods directly impact the effectiveness of CDA in providing targeted interventions and personalized learning strategies in real-world educational contexts. Future research directions are also presented, emphasizing the need for more development of entirely new diagnostic items, hybrid CDMs, and adaptive cognitive diagnostic assessments. Practical recommendations are provided for practitioners, encouraging a strategic approach based on specific assessment goals.

Keywords: *Cognitive diagnostic assessment, cognitive diagnostic models, cognitive diagnostic approaches, language proficiency evaluation, language assessment*

1. Introduction

The compelling benefits of cognitive diagnostic assessment (CDA) have enticed educational systems to develop methodologies that can delve into diversified aspects of students' cognitive abilities. CDA was first introduced in the mid-1980s, combining cognitive psychology's focus on examining the mind through mental representations and processes underlying observable behavior with psychometrics models (Sternberg, 1984). This fusion has attracted researchers and practitioners due to its promising potential (Leighton & Gierl, 2007). Despite being introduced more than three decades ago, diagnostic classification models (DCM) have not seen widespread implementation in educational systems for their intended purposes (Ravand & Baghaei, 2020). Instead, much of the research on CDA has focused on methodological aspects, such as model development and refinement, or retrofitting them to existing non-diagnostic tests (Maas et al., 2024). Meanwhile, CDA was only applied to language assessment in the late 90-s (Buck et al. 1997; Kasai, 1997) and gradually gained recognition in the late 2000s (Lee et al., 2009; Lee & Sawaki, 2009).

CDA has been a focal point of attention due to its capacity to gauge students' capabilities beyond the limitations posed by classical traditional test theory and item response theory (Sessoms & Henson, 2018). In stark contrast to other testing approaches that measure latent traits based on unidimensional or multidimensional constructs, CDA stands out by providing intricate insights into students' micro-skills within the assessed constructs (Ravand & Baghaei, 2020). For example, in assessing candidates' reading skills, CDA is able to extract their abilities in identifying main ideas, understanding new words, and making inferences. This article not only serves to illuminate the historical trajectory and impact of CDA within language proficiency evaluation but also offers practical guidance for educators and researchers. By understanding how CDA has been implemented in the field, readers can glean valuable insights into its application in the development of new assessment items. Through the exploration of CDA's concepts, processes and ongoing debating issues, educators can gain a deeper understanding of how to harness its potential to create more effective assessment tools tailored to the diverse

needs of students. Moreover, by examining the evolution of CDA over time, this article provides a roadmap for educators looking to integrate this innovative approach into their assessment practices, ultimately enhancing their ability to accurately measure and support student learning outcomes. This report will therefore delve into the portrayal of CDA in academic studies, a thorough discussion on the conundrum and also some consideration to better apply CDA in the evaluation of language proficiency.

2. How Can We Understand Cognitive Diagnostic Assessment?

CDA is a specialized way of evaluating students' skills in education. It goes beyond traditional tests that give a single overall score and aims to reveal specific strengths and weaknesses in different thinking areas (Wang et al., 2021). CDA is instrumental in language proficiency evaluation as it provides educators and learners with essential information about the specific areas of language that require attention (Mei & Chen, 2022). CDA helps pinpoint a learner's language abilities in detail when applied to language proficiency evaluation. It identifies which specific language skills a learner has mastered and which ones need more work (Toprak & Çakir, 2018). This diagnostic insight enables the customization of educational programs and pedagogical strategies to cater to individual learning needs. By identifying cognitive strengths and weaknesses, CDA facilitates targeted instruction and intervention, ultimately enhancing language learning outcomes (Rupp et al., 2010). For example, in assessing second language (L2) reading comprehension, CDA can determine if a student struggles with vocabulary, grammar, making inferences, or other reading micro-skills (Shahmirzadi & Marashi, 2023).

The role of CDA in language proficiency evaluation is crucial because it provides detailed insights into the particular aspects of language that need focused instruction or scaffolding (Mei & Chen, 2022). Educators and learners benefit from this diagnostic information, enabling them to tailor educational programs and teaching strategies to address the unique needs of each student. This personalized approach ultimately leads to more effective language learning and teaching outcomes. The process of CDA involves defining cognitive attributes, constructing items, creating a Q-matrix (which links test items to the attributes they measure), and employing cognitive diagnostic models (CDMs) to analyze data (Zhang et al., 2023). This approach allows for a detailed examination of language skills, which proves particularly valuable in large-scale language assessments where precise diagnostic feedback supports effective language teaching and learning initiatives.

Cognitive attributes refer to the specific cognitive skills, knowledge, or problem-solving strategies that students require to complete a particular test task (Li et al., 2021). These attributes are synonymous with sub-skills or micro-skills within the context of CDA. The identification and understanding of these attributes are crucial for creating a Q-matrix, which is an association matrix that describes the relationship between test items and the cognitive attributes they assess (Wang et al., 2021). This Q-matrix also serves as a bridge between the answers provided by students and their mastery patterns of the attributes (Zhang et al., 2023). In retrofitting studies, the attributes are extracted from the existing items, while in studies that develop new cognitive diagnostic items, attributes are used to guide the construction of items.

CDA studies focus primarily on data analysis procedures employing cognitive diagnostic models (CDMs). These models present clear advantages compared to classical test theory (CTT) and item response theory (IRT) in educational assessment (Meng et al., 2023). In contrast to CTT and IRT, CDMs furnish in-depth diagnostic information by deconstructing test-takers performances into specific cognitive attributes or sub-skills. This detailed breakdown allows for a thorough comprehension of individual strengths and weaknesses, paving the way for targeted instruction and personalized learning plans (Ravand & Robitzsch, 2018). CDMs prove especially advantageous in evaluating intricate skills like language proficiency, capturing the diverse components contributing to overall performance. Furthermore, CDMs facilitate the tailoring of assessments to pinpoint specific cognitive processes, providing flexibility in evaluating multidimensional constructs (Li et al., 2021). These models also explicitly delineate cognitive processes, offering a profound understanding of how individuals approach diverse tasks. On top of that, CDMs generate comprehensive diagnostic feedback, guiding educators and learners toward precise interventions and instructional strategies (Toprak & Cakir, 2021).

3. Conundrum and Consideration in Cognitive Diagnostic Assessment

Due to its extensive application, CDA has sparked debates among scholars and researchers, particularly

regarding the selection of models capable of comprehensively analyzing the collected data. The literature extensively documents hundreds of cognitive diagnostic models that have been employed to unravel the complexity of various human skills. Also, scholars engage in ongoing discussions about whether to justify the use of a post-hoc study design in retrofitting existing examination questions or to opt for developing newly designed items from scratch. Additionally, during the construction and validation of the Q-matrices, arguments have arisen regarding the best approach to produce valid and reliable matrices.

In the competition between cognitive diagnostic models, who wins?

Cognitive diagnostic models (CDMs) are a class of psychometric models used in educational assessment to provide detailed information about an individual's specific strengths and weaknesses in various cognitive skills or attributes. Unlike traditional assessment models that provide an overall score, CDMs aim to identify the specific cognitive skills or attributes that an individual has mastered or has not mastered (Ketabi et al., 2021). These models are particularly useful in educational settings as they can provide valuable insights into a student's learning needs and inform targeted instructional strategies (Liao et al., 2024). CDMs are based on the assumption that an individual's performance on a test is influenced by their mastery of a set of underlying skills or attributes, and the models aim to infer the individual's skill mastery profile based on their test responses (Li & Hunter, 2015). Various types of CDMs have been developed, including the General Diagnostic Model (GDM), Fusion Models, Latent Class Analysis (LCA), Deterministic Inputs, Noisy "Or" Gate (DINO), the Additive Cognitive Diagnostic Model (ACDM), log-linear cognitive diagnostic model (ACDM) and many more. These models differ in their underlying statistical and computational approaches, but they all share the common goal of providing detailed diagnostic information about an individual's cognitive skills (Eren et al., 2023; Javidanmehr & Sarab, 2017).

Table 1: Examples of compensatory and non-compensatory models

Compensatory CDM	Non-compensatory CDM
Generalized Deterministic Inputs, Noisy "And" Gate model, G-DINA (de la Torre, 2011)	Deterministic Inputs, Noisy And Gate, DINA (Junker & Sijtsma, 2001)
Additive Cognitive Diagnostic Model, ACDM (de la Torre, 2011)	Reparameterized Unified Model, RUM (DiBello et al., 1995)
Deterministic Inputs, Noisy "Or" Gate Model, DINO (Templin & Henson, 2006)	Reduced Rule Space Model, RRUM (Hartz, 2002)
Linear Logistic Test Model, LLTM (Fischer, 1973)	Long-DINA (Zhan et al., 2019)
	Log-linear Cognitive Diagnostic Model, LCDM (Henson et al., 2009)

The ongoing discourse on cognitive diagnostic models revolves around the pivotal distinction between compensatory and non-compensatory models. Compensatory and non-compensatory models are two types of cognitive diagnostic models (CDMs) used in educational assessment to understand how individuals perform on tests based on their underlying cognitive skills or attributes (Li & Hunter, 2015). The key difference between these models lies in how they account for the relationship between these cognitive skills when making inferences about an individual's performance (Ravand, 2016). In compensatory models, also called disjunctive models, it is assumed that mastery of one cognitive attribute can compensate for the lack of mastery of another attribute (Mohammed et al., 2023). In other words, if an individual is strong in one attribute, it can make up for weaknesses in another attribute when answering test items. This means that in compensatory models, individuals can still perform well on a test even if they have weaknesses in certain attributes, as long as their strengths in other attributes compensate for those weaknesses. Examples of compensatory models include the Deterministic Inputs, Noisy "Or" Gate, DINO (Templin & Henson 2006), the Additive Cognitive Diagnostic Model, ACDM (de la Torre, 2011), and the Generalized Deterministic Inputs, Noisy "And" Gate model, G-DINA (de la Torre, 2011).

On the other hand, it is assumed that all required attributes must be mastered to correctly answer a test item in non-compensatory models, also known as conjunctive models (Tabatabaee-Yazdi & Samir, 2023). It means that weaknesses in any one attribute cannot be compensated for by strengths in other attributes (Effatpanah & Baghaei, 2019). In other words, non-compensatory models are more stringent in their assessment, requiring individuals to demonstrate mastery of all relevant attributes for each test item. Some examples of non-

compensatory models include the Deterministic Input, Noisy "And" gate (DINA) model (Junker & Sijtsma, 2001), Reparameterized Unified Model (RUM) (DiBello et al., 1995), and Reduced Rule Space Model (RRUM) (Hartz, 2022). The main difference between compensatory and non-compensatory models lies in how they account for the relationship between cognitive attributes when making inferences about an individual's performance on a test. Compensatory models allow for strengths in certain attributes to compensate for weaknesses in others, while non-compensatory models require mastery of all relevant attributes for successful performance.

Another different type of CDM classification is saturated models which depends on the number of parameters used to fit the model to the data. A saturated model has enough parameters to perfectly fit the observed data, resulting in a perfect fit with no unexplained variability (Terzi & Sen, 2019). This means that the model is flexible and can perfectly predict the response patterns of individuals based on their mastery or non-mastery of specific cognitive skills (Min et al., 2022). Saturated models also possess the capacity to deal with both compensatory and non-compensatory relationships simultaneously (Dong et al., 2022). However, saturated models may be overly complex and need a large sample size (Sen & Cohen, 2021). Examples of saturated models include the general diagnostic model, GDM (von Davier & Lee, 2019), and the hierarchical diagnostic classification model, HCDM (Templin & Bradshaw, 2013).

Newly-developed CDA items vs retrofitting non-diagnostic items, which is preferable?

Another ongoing discussion in CDA is whether to retrofit an existing item test or develop new measurement items. Both approaches offer different advantages and use distinct ways of conducting studies. Retrofitting studies refer to the process of applying cognitive diagnostic models (CDMs) to existing non-diagnostic tests, to extract diagnostic information from the test results (Mirzaei et al., 2020). This approach involves constructing a Q-matrix, which maps the test items to the cognitive skills they measure, and then fitting a CDM to the test data to estimate the mastery of each skill for each test taker (Toprak & Çakir, 2018). Retrofitting studies are often used when it is not feasible or practical to develop a new diagnostic test from scratch, and have been applied to a variety of high-stakes proficiency exams in language testing. In the context of language assessment, studies have been done to retrofit high-stakes examinations such as the Program for International Student Assessment (PISA) (Chen & Chen, 2015, 2016), Michigan English Language Assessment Battery (MELAB) (Li & Hunter, 2015; Li & Suen, 2013), Test of English as a Foreign Language (TOEFL) (Safari & Ahmadi, 2023; Yi, 2016), College English Test (CET) (Meng et al., 2023; Meng & Fu, 2023; Shi et al., 2024), International English Language Testing System (IELTS) (Mirzaei et al., 2020; Panahi & Mohebbi, 2022) and Progress in International Reading Literacy Study (PIRLS) (Thi & Loye, 2019).

In contrast to retrofitting studies, a distinct approach in cognitive diagnostic assessments involves the creation of new items grounded in the initially specified attributes. These studies undergo various phases before the analysis through cognitive diagnostic models (CDMs) is initiated. The process typically commences with the identification of attributes for measuring selected domains, employing expert judgment, document analysis, and literature review (Toprak & Cakir, 2020). Subsequently, items are meticulously crafted based on these attributes, each assigned a specific number. The next step entails constructing a Q-matrix to delineate the tentative relationship between attributes and items before collecting data from the targeted population (Nallasamy & Khairani, 2022). Ultimately, CDMs are employed to empirically validate the Q-matrix and analyze the data, culminating in the generation of students' mastery profiles (Alavi & Ranjbaran, 2018). Numerous studies on developing new CDA items, conducted across diverse regions and targeting various domains in language assessment, have been previously documented. Among these, a recurrent focus has been on reading comprehension items, chosen by several authors to discern students' proficiency in reading sub-skills (Doe, 2014; Y. Li et al., 2021; Nallasamy & Khairani, 2022; Ranjbaran & Alavi, 2017; Toprak & Cakir, 2021). Another cluster of research endeavors has delved into the creation and validation of cognitive diagnostic items specifically tailored for assessing writing skills (Kim, 2019; Safari & Ahmadi, 2023; Shi et al., 2024). Additional studies have directed their attention toward crafting items gauging proficiency in speaking (Poolsawad et al., 2015) and grammar (Clark & Endres, 2021; Mizumoto & Webb, 2017)

Each approach serves different purposes tailoring to the objectives of the studies. Retrofitting studies follow a different path, where attributes are delineated within the scope of what existing items already assess. Although this approach benefits from the availability of pre-existing items, facilitating a more straightforward Q-matrix

construction, it is limited by constrained attribute specification. The retrofitting process is confined to what the existing items were originally designed to measure. In contrast, developing new CDA items from scratch presents distinct advantages. This is because, in this approach, attributes are first precisely specified, and items are subsequently developed based on these identified attributes. This method allows for greater flexibility in refining and adjusting item quality to accurately measure the specified attributes. Furthermore, the construction of the Q-matrix in this approach is highly tailored to the goals of the studies or the assessment system, ensuring a targeted measurement of candidate skills. While both approaches have their merits, the development of new CDA items stands out for its ability to offer a more customizable and precise measurement of attributes, enhancing the overall quality of the assessment process.

Expert-based or empirical validation, which is better?

A vital component in CDM is the construction of a Q-matrix to map the relationship or association between items and attributes introduced by Tatsuoka (1983). The matrix is typically constructed after attributes have been specified and items have been developed and undergo an iterative process. To validate the relationship between items and attributes, the matrix needs to be validated either through panel judgment or empirical analysis. Some studies have validated the matrix qualitatively using the judgment made by expert panels (Liu et al., 2017; Ravand, 2016). The process of selecting experts typically involves qualitative methods, while their consensus is often quantified, for example, through the utilization of Fleiss' Kappa to establish a shared matrix (Shi et al., 2024). Another method of quantitatively validating these matrices through expert judgment is Interpretive Structural Modeling (ISM) as shown in a study by Zhang et al. (2024).

However, some scholars have expostulated the caliber of qualitative judgment and put forward empirical methods to validate the matrix. Recent advancements have proposed empirical methods using quantitative approaches for Q-matrix validation (Chen et al., 2015; de la Torre & Chiu, 2016; DeCarlo, 2012; Desmarais & Naceur, 2013). Some of these approaches are entirely data-driven, with underlying attributes derived from test takers' responses (Meng et al., 2023). Others are designed to identify potential misspecifications in expert-defined provisional Q-matrices, suitable for situations where misspecifications can be identified (DeCarlo, 2012; Templin & Henson, 2006).

The battle between different methods to validate the Q-matrix reflects a dynamic landscape in the field of cognitive diagnostic modeling (CDM). Researchers and practitioners grapple with choosing the most effective approach among the array of available methods. On one front, the factorization method proposed by Desmarais and Naceur (2013) Emphasizes the iterative refinement of an expert-defined Q-matrix based on test takers' responses. On another front, the Bayesian Extension introduced by DeCarlo (2012) introduces a probabilistic approach to acknowledge uncertainty in the Q-matrix. Meanwhile, the general method of empirical Q-matrix validation, developed by de la Torre and Chiu (2016), offers a comprehensive solution compatible with the G-DINA model and specific DCMs. In contrast, the regularized latent class analysis (RLCA) method, proposed by Chen et al. (2018), presents a non-provisional approach, deriving the Q-matrix directly from test responses. The battle extends to the sphere of model comparison and fit indices, where metrics like the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are employed to assess the appropriateness of different Q-matrices. Additionally, the indices for classification consistency and accuracy, as presented by Cui (2012), add another layer to the skirmish, contributing to the ongoing discourse on the reliability and validity of classifications made by DCMs. In this dynamic arena, the quest for the most robust and applicable method for Q-matrix validation persists, shaping the trajectory of CDM research and application.

4. Future Directions and Recommendations

As the field of cognitive diagnostic assessment (CDA) continues to evolve, researchers are faced with the challenges of exploring and refining existing cognitive diagnostic models (CDMs). The current landscape is dominated by compensatory and non-compensatory models, each with its strengths and limitations. Future research endeavors could focus on expanding the repertoire of Cognitive Diagnostic Models (CDMs) by developing hybrid models that integrate the advantages of both compensatory and non-compensatory approaches, especially in the context of language proficiency evaluation. This hybridization may provide a more significant understanding of how cognitive skills interact and influence overall performance, offering a comprehensive assessment framework, particularly in complex skills in language acquisition. Additionally, the

exploration of novel CDMs with enhanced computational approaches and statistical foundations can contribute to the refinement of diagnostic accuracy and reliability. The quest for the most effective CDM remains an ongoing journey, opening avenues for researchers to innovate and enhance the precision of cognitive diagnostic assessment. On top of that, practitioners engaging in cognitive diagnostic assessment (CDA) should carefully consider the implications of choosing between compensatory and non-compensatory models. The decision has profound consequences for the depth of diagnostic insights provided. While compensatory models allow for a degree of flexibility by acknowledging that strength in one skill can compensate for weaknesses in another, non-compensatory models demand a more stringent mastery of all relevant skills for successful performance. Practitioners should align their choice with the specific objectives of the assessment and the educational context in which it is applied.

Additionally, the ongoing debate on retrofitting existing items versus developing new diagnostic items calls for a strategic approach. When retrofitting is deemed appropriate, practitioners must ensure the alignment of existing items with the intended cognitive attributes. Conversely, when developing new items, the careful crafting of items based on identified attributes is crucial for the validity and reliability of the diagnostic process. Future studies should emphasize the development of entirely new items rather than relying solely on retrofitting existing non-diagnostic tests. This shift in focus is crucial for the continuous evolution of CDA, ensuring that the assessment tools are aligned with the ever-changing landscape of language proficiency evaluation. By delving into the creation of innovative diagnostic items, researchers can address the limitations of retrofitting studies and contribute to the refinement and expansion of the CDA framework.

Apart from that, the clash between expert-based and empirical validation methods for Q-matrix in cognitive diagnostic assessment (CDA) presents a dynamic arena for future research. Researchers and practitioners should focus on exploring the strengths and limitations of different validation methods to enhance the reliability and validity of Q-matrices. Comparative studies that systematically evaluate the performance of various validation approaches, such as the factorization method, Bayesian Extension, regularized latent class analysis (RLCA), and model comparison metrics, will contribute to a more profound understanding of their applicability in different contexts. Practical guidelines for practitioners in choosing the most suitable validation method based on their specific assessment goals and constraints should also be a focus of future research endeavors.

The integration of adaptive cognitive diagnostic assessment represents a promising avenue for future research. Adaptive CDA tailors the assessment process in real-time based on the test taker's responses, dynamically adjusting the difficulty and content of subsequent items. This adaptive approach has the potential to enhance the efficiency and precision of diagnostic assessments by focusing on the specific cognitive skills relevant to an individual's proficiency level. Researchers should explore adaptive strategies within the CDA framework and investigate their implications for improving the accuracy of diagnostic feedback and the overall effectiveness of language learning interventions. Additionally, CDA holds the capacity to assist teachers and educational practitioners in discerning students' mastery of micro-skills within the classroom. This can be particularly valuable for formative assessments or in-class evaluations. Consequently, beyond research concentrated on high-stakes international assessment systems, a strategic initiative should be undertaken to seamlessly integrate CDA into low-stakes settings.

Furthermore, while the adoption of CDA has been notably successful in specific nations like the United States of America, Iran, and China, there is a compelling case for broader global participation. Other countries especially Asian countries would greatly enhance their assessment systems by actively integrating CDA methodologies. CDA possesses the capability to enhance the quality of assessment systems, whether in high-stakes or low-stakes contexts, thereby indirectly elevating teacher practices and the overall education system.

5. Conclusion

In the exploration of cognitive diagnostic assessment for language proficiency evaluation, this article has delved into the wide-ranging possibilities of assessing students' cognitive skills, moving beyond traditional testing methodologies. From the foundational understanding of CDA's emergence to the ongoing conundrum of different classifications of models, and the strategic considerations in developing new diagnostic items, the discussion on

this discourse has illuminated the significance of evolving assessment approaches. With its ability to unravel the specific strengths and weaknesses of individuals in various cognitive domains of language skills, CDA stands as a beacon for educational systems aiming to tailor instruction to the unique needs of learners. However, CDA is not without its challenges, and practitioners must remain vigilant to potential threats. Given its capability to offer detailed insights into an individual's abilities, the assessment demands a meticulous approach to rating and score analysis. This process requires practitioners to navigate issues such as cultural biases in assessment tools and the potential impact of test anxiety on accurate measurements. Additionally, considering environmental factors and addressing motivational aspects becomes crucial to ensuring the reliability and validity of the intricate information CDA aims to provide. In conclusion, the rigorous research on the methodological aspects of CDA should be coupled with initiatives to develop new CDA items as empirical evidence of how the application of CDA can be meaningful and impactful. With the dynamic nature of language proficiency evaluation, CDA stands out as a great method to capture the intricacies of cognitive processes in language acquisition.

Funding: This study was funded by Ganjaran Penerbitan GP-KO21854 and TAP-KO21854 Universiti Kebangsaan Malaysia.

Competing Interests: The authors declare that they have no competing interests.

Acknowledgment: All authors contributed equally to the conception and design of the study.

References

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.

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

Chen, H. T., Pan, H. L. W., Morosanu, L., & Turner, N. (2018). Using logic models and the action model/change model schema in planning the learning community program: A comparative case study. *Canadian Journal of Program Evaluation*, *33*(1), 49–68. https://doi.org/10.3138/cjpe.42116

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix-based diagnostic classification models. *Journal of the American Statistical Association*, *110*(510), 850–866. https://doi.org/10.1080/01621459.2014.934827

Clark, T., & Endres, H. (2021). Computer-based diagnostic assessment of high school students' grammar skills with automated feedback–an international trial. *Assessment in Education: Principles, Policy and Practice*, *28*(5–6), 602–632. https://doi.org/10.1080/0969594X.2021.1970513

Cui, Y. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*(1), 19–38.

de la Torre, J. (2011). The Generalized DINA Model framework. *Psychometrika*, *76*(3), 510–510. https://doi.org/10.1007/s11336-011-9214-8

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*(2), 253–273. https://doi.org/10.1007/s11336-015-9467-8

DiBello, L. V., Stout, W. F., & Roussos, L. (1995). Unified cognitive psychometric sessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.). Cognitively diagnostic assessment (pp. 361–390). Hillsdale, NJ: Erlbaum.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian Extension of the DINA Model. *Applied Psychological Measurement*, *36*(6), 447–468. https://doi.org/10.1177/0146621612449069

Desmarais, M. C., & Naceur, R. (2013). A matrix factorization method for mapping items to skills and for enhancing expert-based Q-matrices. *Artificial Intelligence in Education: 16th International Conference*, 441–450.

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

Dong, M., Gan, C., Zheng, Y., & Yang, R. (2022). Research trends and development patterns in Language Testing

over the past three decades: A bibliometric study. *Frontiers in Psychology*, *13*, 1–15. https://doi.org/10.3389/fpsyg.2022.801604

Effatpanah, F., & Baghaei, P. (2019). Diagnosing EFL learners' writing ability: A diagnostic classification modeling analysis. *Language Testing in Asia*, *9*(12), 1–23.

Eren, B., Gündüz, T., & Tan, Ş. (2023). Comparison of methods used in the detection of DIF in Cognitive Diagnostic Models with traditional methods: Applications in TIMSS 2011. *Journal of Measurement and Evaluation in Education and Psychology*, *14*(1), 76–94. https://doi.org/10.21031/epod.1218144

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359–374.

Hartz, S. M. (2002). A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality. Dissertation Abstracts International: Section B: The Sciences and Engineering, 63(2-B), 864.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210. https://doi.org/10.1007/s11336-008-9089-5

Javidanmehr, Z., & Sarab, M. R. A. (2017). Cognitive diagnostic assessment: Issues and considerations. *International Journal of Language Testing*, *7*(2), 73–98.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272. https://doi.org/10.1177/01466210122032064

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.

Kim, Y. (2019). Developing and validating empirically derived diagnostic descriptors in ESL academic writing. *The Journal of Asia TEFL*, *16*(3), 906–926.

Lee, Y., Sawaki, Y., & Lee, Y. (2009). Application of three Cognitive Diagnosis Models to ESL reading and listening assessments. *Language Assessment Quarterly ISSN: 6*(3), 239–263. https://doi.org/10.1080/15434300903079562

Lee, Y. W., & Sawaki, Y. (2009). 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. (2007). *Cognitive Diagnostic Assessment for Education*. Cambridge University Press.

Li, H., & Hunter, C. V. (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*(September), 1–15. https://doi.org/10.3389/fpsyg.2021.714568

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*(December), 1–13. https://doi.org/10.3389/fpsyg.2021.786612

Liao, M., Jiao, H., & He, Q. (2024). Explanatory Cognitive Diagnosis Models Incorporating Item Features. *Journal of Intelligence*, *12*(3), 32. https://doi.org/10.3390/jintelligence12030032

Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2017). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and Psychological Measurement*, *77*(2), 220–240. https://doi.org/10.1177/0013164416645636

Maas, L., Madison, M. J., & Brinkhuis, M. J. S. (2024). Properties and performance of the one-parameter log-linear cognitive diagnosis model. *Frontiers in Education*, *9*. https://doi.org/10.3389/feduc.2024.1287279

Mei, H., & Chen, H. (2022). Cognitive diagnosis in language assessment: A thematic review. *RELC Journal*, 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

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

Mizumoto, A., & Webb, S. A. (2017). Developing and evaluating a computerized adaptive testing version of the Word Part Levels Test. *Language Testing*, *36*(1), 1–23. https://doi.org/10.1177/0265532217725776

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.

Nallasamy, R., & Khairani, A. Z. Bin. (2022). Development and validation of reading comprehension assessments by using the GDINA Model. *Malaysian Journal of Social Sciences and Humanities (MJSSH)*, *7*(2), 1–13. https://doi.org/10.47405/mjssh.v7i2.1278

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

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

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. (2020). 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. In T. D. Little (Ed.), *Measurement: Interdisciplinary Research and Perspectives* (Issue 1). The Guilford Press. https://doi.org/10.1080/15366367.2018.1434349

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

Sen, S., & Cohen, A. S. (2021). Sample size requirements for applying Diagnostic Classification Models. *Frontiers in Psychology*, *11*, 1–16. https://doi.org/10.3389/fpsyg.2020.621251

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., & 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

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

Tabatabaee-yazdi, M., & Samir, A. (2023). On the identifiability of Cognitive Diagnostic Models: Diagnosing students' translation ability. *Journal of Language & Education*, *9*(1), 138–157.

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.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of Diagnostic Classification Modelexaminee estimates. *Journal of Classification*, *30*, 251–275. https://doi.org/10.1007/s00357-013

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using Cognitive Diagnosis Models. *Psychological Methods*, *11*(3).

Terzi, R., & Sen, S. (2019). A non-diagnostic assessment for diagnostic purposes: Q-matrix validation and Item-Based Model git evaluation for the TIMSS 2011 Assessment. *SAGE Open*, *9*(1), 1–11. https://doi.org/10.1177/2158244019832684

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.

Toprak, T. E., & Cakir, A. (2020). 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

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

Toprak, T. E., & Çakir, A. (2018). Where the rivers merge: Cognitive diagnostic approaches to educational assessment. *Kuramsal Eğitimbilim*, *11*(2), 244–260. https://doi.org/10.30831/akukeg.363915

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

Wang, D., Cai, Y., & Tu, D. (2021). Q-matrix estimation methods for Cognitive Diagnosis Models: Based on Partial Known Q-Matrix. *Multivariate Behavioral Research*, *56*(3), 514–526. https://doi.org/10.1080/00273171.2020.1746901

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

Zhan, P., Jiao, H., Liao, D., & Li, F. (2019). A longitudinal higher-order Diagnostic Classification Model. *Journal of Educational and Behavioral Statistics*, *44*(3), 251–281. https://doi.org/10.3102/1076998619827593

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

Zhang, S., Liu, J., & Ying, Z. (2023). Statistical applications to cognitive diagnostic testing. *Annual Review of Statistics and Its Application*, *10*, 651–678.