Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 15, No. 2; 2025, October

ISSN 2476-5880
https://www.ijlt.ir

An Exploratory Mixed Methods Study of Grammatical Range and Accuracy in IELTS: A Genuinely Diagnostic Approach to Cognitive Diagnostic Assessment

Bahar Barghi[1], Jafar Afshinfar [2]*, Seyyed Mohammad Alavi[3], Manoochehr Jafarigohar[4]*, Hassan Soleimani[5]

ARTICLE INFO

Article History:
Received: October 2024
Accepted: December 2024

KEYWORDS
Cognitive Diagnostic Assessment,
Grammatical Range and Accuracy,
IELTS,
Language Assessment

ABSTRACT

In the second language (L2) assessment realm, cognitive diagnostic assessment (CDA) emerges as an exceptionally practicable methodology, enabling a meticulous analysis of linguistic competencies and providing detailed insights into learners' proficiencies and deficiencies, thereby charting precise remedial pathways. The primary objective of this research was to develop a cognitive model of attributes underlying an International English Language Testing System (IELTS) assessment descriptor, namely grammatical range and accuracy. The model sought to enable the development of a diagnostic instrument informed by CDA, which would investigate candidates' grammatical strengths and weaknesses to improve their performance in the IELTS. Through a multi-stage process involving qualitative data collection, interpretation, and synthesis, a comprehensive scheme emerged, categorizing cognitive attributes into two main areas: (a) knowledge of grammatical forms, including pronouns, determiners/quantifiers, adjectives, adverbials, nouns/noun phrases, verbs/tenses, and prepositions; and (b) familiarity with structural nuances, including punctuation and structural sophistication. These nine micro-level attributes comprised several sub-components aligned with three proficiency classes: A1-A2, B1-B2, and C1-C2. The model laid the groundwork for developing a three-booklet multiple-choice test. Alongside conducting pilot testing and item analysis, the study employed a saturated psychometric model to validate the CDA-informed instrument. The results confirmed the instruments' internal consistency, validity, satisfactory fit, and effectiveness in classifying examinees based on attribute-specific mastery levels. The findings underscored severe weaknesses in punctuation, structural complexity, and verb tense usage, pinpointing crucial areas demanding targeted instructional enhancement. The theoretical implications highlight a refined understanding of grammatical competencies, while pedagogically, the results advocate for targeted teaching strategies.

[1] Department of English, Faculty of Literature and Foreign Languages, Payame Noor University, Tehran, Iran Email: bahar.barghi@student.pnu.ac.ir
[2] Department of English, Faculty of Literature and Foreign Languages, Payame Noor University, Tehran, Iran Email: ja.afshinfar@pnu.ac.ir
[3] Department of English, Faculty of Literature and Foreign Languages, University of Tehran, Tehran, Iran Email: smalavi@ut.ac.ir
[4] Department of English, Faculty of Literature and Foreign Languages, Payame Noor University, Tehran, Iran Email: Jafari@pnu.ac.ir
[5] Department of English, Faculty of Literature and Foreign Languages, Payame Noor University, Tehran, Iran Email: h_soleimanis@pnu.ac.ir

Cite this paper as: Barghi , B., Afshinfar, J., Alavi, M, Jafarigohar, M, & Soleimani, H. (2025). An exploratory mixed methods study of grammatical range and accuracy in IELTS: A genuinely diagnostic approach to cognitive diagnostic assessment. *International Journal of Language Testing*, *15*(2), 19–42. https://doi.org/10.22034/ijlt.2024.481434.1374

1. Introduction

Grammatical knowledge constitutes a fundamental cornerstone of language proficiency (Ellis, 2006), particularly within foreign language (FL) contexts such as Iran, where limited exposure to authentic language and reliance on formal instruction accentuate its significance (Rahimi & Zhang, 2015). In such environments, grammar not only facilitates the construction of grammatically accurate sentences but also underpins comprehensive communicative competence (Larsen-Freeman & Celce-Murcia, 2015). For Iranian learners, where English often serves as a gateway to academic and professional success, mastery of grammar is indispensable for attaining higher levels of language proficiency (Kargar Behbahani et al., 2024). Grammar provides the essential structural framework for both receptive (listening and reading) and productive (speaking and writing) skills, thereby rendering it crucial for effective communication and fluency (Ruzmetova, 2024).

In light of grammar's vital role in language acquisition and performance, diagnostic approaches to grammar instruction have gained significant traction, enabling educators to assess specific aspects of grammatical competence and deliver targeted, individualized feedback. These diagnostic assessment models aim to identify learners' strengths and pinpoint areas of difficulty, a process particularly valuable in FL contexts like Iran, where instructional time is limited and efficient, and goal-oriented learning is of utmost priority (Ketabi et al., 2021; Pishghadam & Miri, 2021). This approach aligns with CDA models, which provide a structured framework for identifying learners' knowledge gaps and planning effective, targeted interventions (Lee & Sawaki, 2009).

Unlike traditional assessments that often focus on overall scores or rankings, CDA models prioritize analyzing specific skill areas to uncover detailed insights into each learner's unique language profile (Jang, 2009). As for grammatical competence assessment, CDA can reveal whether learners struggle with verb tenses, subject-verb agreement, or complex sentence structures, allowing instructors to tailor instruction accordingly and address these areas in a time-sensitive manner (Li et al., 2021). By aligning instructional methods with learners' precise needs, CDA supports the development of grammatical accuracy and communicative competence, enhancing fluency and expressive ability even in highly structured, exam-oriented educational systems (Zhang & Thompson, 2004)

The application of diagnostic assessment in grammar is particularly pertinent for high-stakes language exams like the IELTS, widely recognized in Iran and other FL contexts as a benchmark for academic and professional proficiency (Pilcher & Richards, 2017). The IELTS assesses productive skills in speaking and writing, where grammatical range and accuracy are essential scoring criteria. However, the exam does not explicitly test grammatical sub-skills, placing implicit pressure on candidates to demonstrate accurate and varied grammatical structures without receiving targeted feedback on their strengths or weaknesses in grammar. Consequently, an effective diagnostic model that assesses grammatical ability for high-stakes exams can provide invaluable insights, empowering learners to refine their grammar skills to meet test requirements and enhance their overall communicative competence (Shintani & Ellis, 2018).

Centered on grammatical range and accuracy in IELTS, a pivotal criterion for evaluating writing and speaking proficiency, this study aimed to propose a cognitive model assessing the foundational grammar skills essential for achieving success in these productive skill assessments. The study acknowledged that the depth of grammatical knowledge correlates directly with writing and speaking scores, influencing performance in the whole test (Pilcher & Richards, 2017). Accordingly, the proposed model aimed to lay the groundwork for enhancing overall IELTS performance. The research pursued developing a purpose-built diagnostic tool and verifying it through psychometric model fit and skill mastery indices aligned with its constructs. The study employed a genuinely diagnostic approach to model and instrument construction to address a significant gap in the literature for domain-specific grammatical constructs. Its implications may extend to informing L2 learning practices, particularly in remedial instruction within IELTS preparation courses. By offering candidates and instructors a structured approach to meticulous IELTS preparation, the detailed feedback provided by the CDA-informed test could enrich both intensive and standard IELTS instructional curricula. The following inquiries guided the focused trajectory of the research.

1. What cognitive framework can delineate the essential attributes for demonstrating proficiency in grammatical range and accuracy in IELTS?
2. Can the CDA-informed tool, based on the emerging cognitive model, effectively generate detailed diagnostic insights into the mastery or non-mastery of the tests' attributes and measurement properties?

2. Review of Literature

2.1. A Theoretical Perspective on L2/FL Grammar Development and Assessment

A plethora of models and frameworks have been proposed to illuminate the complex process of grammar acquisition in L2 learning. One foundational theory, Krashen's Input Hypothesis, posits that language acquisition occurs through exposure to comprehensible input that surpasses the current proficiency level (Krashen, 1985). This theory underscores the importance of authentic language exposure and meaningful communication in facilitating natural grammar acquisition. Complementing this, the Interaction Hypothesis highlights the role of interaction in language development, emphasizing the significance of negotiation and feedback in grammatical uptake (Long, 1996). Furthermore, Task-Based Language Teaching (TBLT) has emerged as a pedagogical approach that prioritizes meaningful tasks to engage learners in authentic communication, enhancing both fluency and accuracy. Research suggests that task complexity can positively influence grammatical accuracy, stimulating the use of advanced grammatical structures (Ellis, 2006).

Building upon these foundational theories, the Lexical Approach posits that grammar is most effectively acquired through understanding and using lexical phrases and chunks, improving learners' ability to produce accurate language (Lewis, 1993). Cognitive grammar and usage-based approaches further emphasize the role of language use in shaping grammatical knowledge, highlighting the significance of frequency and exposure in internalizing grammatical structures (Ellis, 2016). Dynamic Systems Theory (DST) offers a more holistic perspective, portraying language development as a complex, adaptive system influenced by cognitive processes and contextual variables. This theory suggests that learners' grammatical development is nonlinear and can be impacted by individual differences and motivation (de Bot et al., 2007). Collectively, the theories and models discussed above provide valuable insights into the mechanisms of grammar acquisition and offer practical implications for language instruction, particularly in exam-oriented contexts.

The foundation of L2 grammar assessment rests upon a comprehensive understanding of the complex processes underlying grammatical competence acquisition and the diverse factors that influence this development. Traditional assessment models often prioritize a holistic assessment of language proficiency, relying on standardized tests that emphasize broad metrics such as accuracy and fluency. However, these approaches may overlook the specific grammatical needs of individual learners (Alderson, 2005), potentially leading to generalized feedback that is less effective in promoting targeted improvement.

Recent advancements in assessment practices highlight the need to align assessments with instructional goals, especially in high-stakes contexts like the IELTS, where grammatical accuracy and range are critical for success. Integrating diagnostic assessment into grammar instruction provides a powerful tool for enhancing grammatical proficiency in such settings. Diagnostic models identify learners' specific strengths and weaknesses, enabling targeted, individualized feedback that addresses their needs (Lee & Sawaki, 2009). This tailored approach fosters a deeper understanding of grammar and empowers learners to develop strategies for improved exam performance. Through CDA frameworks, educators can pinpoint areas for improvement and implement interventions aligned with IELTS demands, ultimately enhancing learners' grammatical range and accuracy in both spoken and written outputs (Shintani & Ellis, 2018).

2.2. Advancements in CDA: Frameworks, Models, and Approaches

The genesis of modern language assessment can be traced to the mid-20th century, a period dominated by two seminal theoretical frameworks: Classical Test Theory (CTT) (Spearman, 1904) and Item Response Theory (IRT) (Lord, 1952). These foundational theories provided the mathematical scaffolding necessary to elucidate the intricate relationship between latent learner traits and their observable manifestations, encompassing proficiency levels, item difficulties, and the multidimensional nature of language ability (Choi et al., 2012). Despite their utility, these traditional approaches faced considerable criticism for their inability to delve into the cognitive processes underlying learner performance, their limited capacity to provide formative feedback, and their potential to induce psychological distress (Rupp et al., 2010).

In response to these shortcomings, the educational community sought to integrate the insights of cognitive psychology with established principles of measurement. This intellectual fusion gave birth to CDA, a paradigm shift that has revolutionized the field of language assessment (Leighton et al., 2004). CDA is characterized by its focus on identifying learners' specific strengths and weaknesses across a range of cognitive skills and knowledge constructs (Lee & Sawaki, 2009).

Two prominent frameworks have shaped the development of diagnostic assessments: Evidence-Centered Design (ECD) (Mislevy, 1994) and Cognitive Design System (CDS) (Embretson, 1998). Both approaches prioritize cognitive considerations in test development, with ECD emphasizing a systematic, multi-layered approach that encompasses domain analysis, cognitive modeling, conceptual framework development, assessment implementation, and delivery (Mislevy et al., 2002). CDS, on the other hand, focuses on the rigorous specification of assessment goals, the development of cognitive models, and the meticulous evaluation of test items (Embretson & Gorin, 2001).

Cognitive Diagnostic Models (CDMs) have emerged as powerful tools for understanding and assessing learner performance. These models, rooted in psychometric theory and cognitive psychology, classify learners into distinct mastery categories based on their response patterns to assessment items (Leighton & Gierl, 2007). Early taxonomies, such as those proposed by Hartz (2002), differentiated between CDMs based on the rule space model (RSM) and the linear logistic trait model (LLTM). Subsequent classifications have expanded to include conjunctive and disjunctive models (Rupp & Templin, 2008), specific and general models (de la Torre, 2011), and compensatory and non-compensatory models (DiBello et al., 2007).

The implementation of CDA has been approached from two primary perspectives: diagnostically constructed designs and non-diagnostically constructed designs. The former involves the creation of assessments specifically designed to measure targeted cognitive attributes, while the latter involves the retrospective inference of cognitive attributes from existing assessments (Jang, 2008). The genuinely diagnostic approach, characterized by its inductive nature, enables the collection of rich, inferential evidence regarding the underlying cognitive skills (Fan et al., 2021). In contrast, the retrofitted approach relies on techniques such as expert judgment and think-aloud protocols to extract cognitive attributes from pre-existing assessments (Jang, 2008).

The literature on diagnostic assessment is replete with studies exploring how CDA tenets could be incorporated into L2 assessment (e.g., Fan et al., 2021; Lee & Sawaki, 2009; von Davier, 2005). While numerous studies have adopted a retrofitted approach to CDA in L2 testing, research on genuinely diagnostic approaches, which involve developing tailored cognitive models and instruments, remains relatively scarce (Williamson, 2023). This scarcity may be attributed to the time-consuming nature of the approach and the lack of established theoretical frameworks specifically designed for L2 assessment. Recent studies have begun to address this gap by developing cognitive models and CDA-informed instruments for L2 reading (e.g., Toprak & Cakir, 2021) and listening comprehension (e.g., Ma & Meng, 2014). However, research on productive skills, such as speaking and writing, and their underlying sub-skills, like vocabulary and grammar, has remained scarce.

2.3. General Deterministic Input, Noisy-and-Gate (G-DINA) Model

Among all CDMs, G-DINA seemed to suit the current study's objectives owing to its saturated, simplistic approach to theorizing inter-attribute relationships (Li et al., 2021). Based on a G-DINA model, the probability of a correct answer is beyond simply adding the success probability of cognitive attributes involved in the test. Instead, the model takes account of the positive and negative effects of between-attribute interactions on providing correct responses to different items within the test. The model's foresight for considering between-attribute associations could account for the widely shared belief that the model is ideally appropriate for assessing language constructs whose underlying attributes are tightly interwoven (Zhao et al., 2020). The G-DINA model groups test-takers into $2^{k_j^*}$classes, in which $k_j^* = \sum_{k=1}^{k} q_{jk}$ represents the number of cognitive attributes needed to provide a correct answer to item j (de la Torre, 2011).

2.4. Empirical Background to the Study

While grammar constitutes a fundamental aspect of language proficiency, its role in diagnostic assessment has been relatively under-explored. Geramipour et al. (2021) investigated the suitability of various cognitive diagnostic models, including the DINA, DINO, and G-DINA models, for predicting the associations between cognitive attributes underlying grammatical competence in a Master's degree entrance examination. Their findings indicated that the G-DINA model provided the best fit for the grammar data. Similarly, Park and Cho (2011) applied CDMs to a 40-item grammar test, identifying learner strengths and weaknesses in areas such as verb use, idiom use, and subject-verb agreement. This study demonstrates the potential of CDMs to provide detailed diagnostic feedback on grammatical knowledge. Yi (2017) conducted a comparative analysis of five compensatory and non-compensatory CDMs (LCDM, C-RUM, DINA, DINO, and NIDO) to determine the most suitable model for diagnostic assessment purposes. Based on various fit indices, the LCDM and C-RUM models emerged as the most appropriate choices for fitting the response data.

3. Method

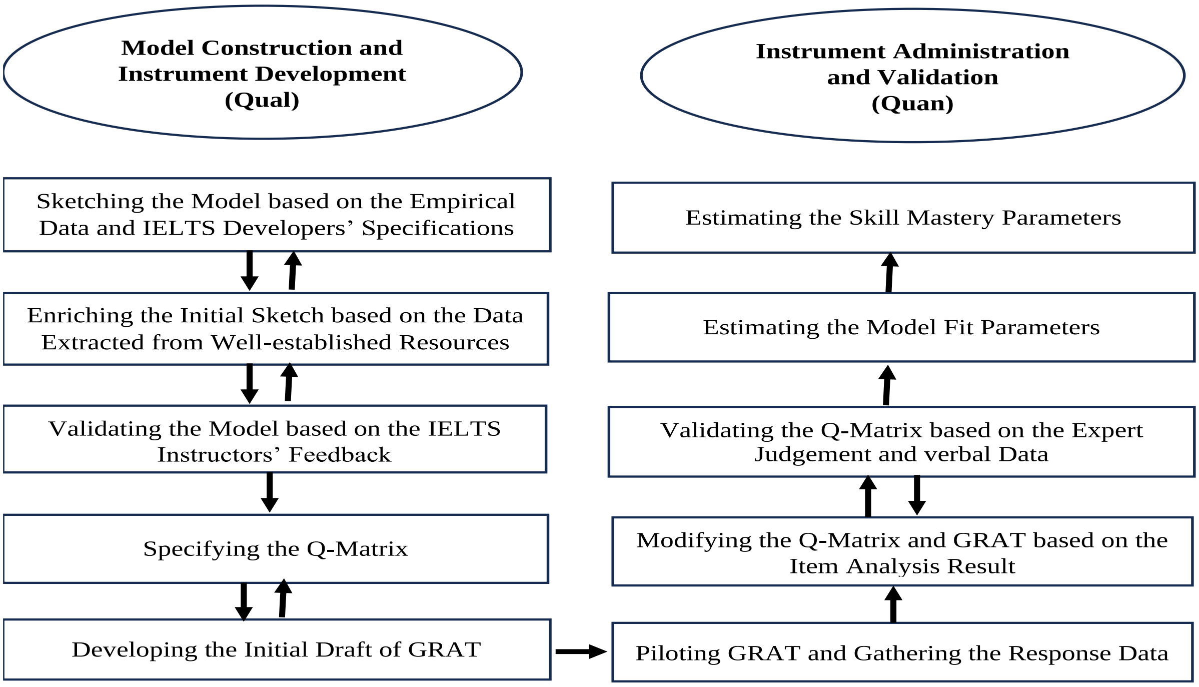
3.1. Conceptual Framework of the Study

The genuinely diagnostic approach conducted in the current study encompassed the multiple procedural steps in Embretson’s (1998) CDS (see the Literature section). The model development process, a primary step in CDS, was mainly grounded on the conceptual underpinnings of cognitive model construction proposed by Leighton and Gierl (2007), which provides a “simplified description of human problem solving on standardized educational tasks, which helps to characterize the knowledge and skills students at different levels of learning have acquired and to facilitate the explanation and prediction of students’ performance” (p. 6). Based on the three-component rubric of cognitive model evaluation proposed by Leighton and Gierl (2007), the emerging model was controlled iteratively for fine-grain size, measurability, and instructional relevance (see Table 1 for more details), dealing with determining cognitive attributes’ depth and breadth, evaluating models’ capability to inform test items suitable for measuring different cognitive attributes, and examining models’ inclusion of attributes instructionally meaningful to a given class of educational stakeholders, respectively.

3.2. Design of the Study

The study’s peculiarities (e.g., qualitative priority, complementarity-developmental purposes, and sequential timing) entailed an exploratory sequential mixed methods (Qual→Quan) design, which commenced with an exploratory qualitative model construction and instrument development phase and moved gradually to a subordinate quantitative phase targeted at pilot administration and validation of the instrument (Creswell & Plano Clark, 2018).

Figure 1
Procedural Steps of the Research Design


Model Construction and Instrument Development (Qual)
Instrument Administration and Validation (Quan)
Sketching the Model based on the Empirical Data and IELTS Developers' Specifications
Enriching the Initial Sketch based on the Data Extracted from Well-established Resources
Validating the Model based on the IELTS Instructors' Feedback
Specifying the Q-Matrix
Developing the Initial Draft of GRAT
Piloting GRAT and Gathering the Response Data
Modifying the Q-Matrix and GRAT based on the Item Analysis Result
Validating the Q-Matrix based on the Expert Judgement and verbal Data
Estimating the Model Fit Parameters
Estimating the Skill Mastery Parameters

As shown in Figure 1, The research encompassed three procedural steps, including an introductory cognitive model development phase, a supplementary instrument development phase, and a concluding instrument validation phase. The first and second phases drew on the existing empirical data, specifications of the testing construct provided by the official enterprises in charge of IELTS, information gathered from well-established IELTS resources, and experienced IELTS instructors' feedback. The instrument, entitled Grammatical Rang and Accuracy Test (GRAT), was intended to measure IELTS candidates' strengths and weaknesses in grammatical constructs underlying productive IELTS modules. On the other hand, the third quantitative instrument validation phase relied upon response data gathered through the pilot administration of GRAT. Based on the research's conceptual framework, each qualitative and quantitative phase encompassed several procedural steps.

3.3. Participants

As entailed by the mixed exploratory design, different groups of participants contributed to the progress of the research phases. The first participant group, chosen through convenience sampling, comprised 19 male (n = 7) and female (n = 12) experienced IELTS instructors who specialized in IELTS preparation courses for ten years or more. As Wu and Thompson (2020) reported, though opening room for sampling bias and unknown errors, convenience sampling is presumed to be a cost and time-effective sampling method for small-scale mixed-methods studies with limited time and funds. The instructors were Ph.D. graduates (n = 5) and undergraduates (n = 14) in English literature (n = 3), translation (n = 3), and teaching (n = 13). The sample was confined to 19 IELTS instructors since information saturation materialized after interviewing the last member of the one-by-one chosen sample. Five of these instructors, enjoying the experience of developing L2 tests, contributed to the progress of the instrument development phase. Based on a snowball sampling method, every member of the instructor panel introduced candidates inclined to participate in a diagnostic test on grammatical structures required for thriving in IELTS. Ultimately, a sample of 584 IELTS candidates, ranging in English proficiency from intermediate (B1) to advanced (C1), constituted the candidate sample. Of all the candidates, 17 advanced-level ones made a voluntary contribution to the pilot administration of the initially developed version of GRAT, and the others (n = 567) took either the pen-and-paper (n =

144) or online version (n = 423) of GRAT (the finalized version). A total of 12 volunteers from the IELTS candidate also attended the in-depth interviews and think-aloud activities required for Q-matrix validation. Table 1 displays the demographics of the participant groups.

Table 1
Demographics of the Participant Groups

Phase	Participant	Age		Gender		N
		M	SD	Male	Female	
Model/Instrument Development	IELTS Instructors	42.7	5.8	7	12	19
Instrument Verification	IELTS candidates	26.7	6.9	8	4	12
GRAT Piloting	IELTS candidates	23.1	6.0	9	8	17
GRAT Verification	IELTS Candidates	28.6	2.1	244	323	567

3.4. Instruments and Materials

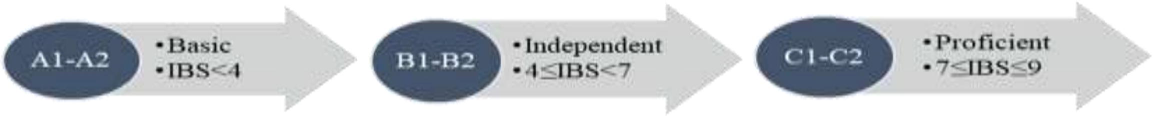
3.4.1. Semi-structured Interviews. The data required to verify the initially built CDA model were gathered by having the IELTS instructors attend semi-structured interviews, including a brief description of the research objectives, a detailed account of the model in progress, and three questions about the fine-grain size, measurability, and instructional relevance of the attributes and their underlying sub-skills in the model under development. They were also asked to give their opinion about the proficiency levels associated with every sub-skill in the model. The Interviews were not time-limited; therefore, the instructors had ample time to propose required amendments to the model based on their first-hand experience of directing IELTS preparation courses.

3.4.2. Think-aloud Protocol. A think-aloud protocol helped the 12-member IELTS candidate group verbalize their thoughts on grammatical structures/skills/competencies while responding to different items of GRAT. The protocol asked the participants not to refuse to say whatever comes into their minds (i.e., what they see, think, feel, and perform) while dealing with every single item in the test. It also invited them to verbalize their thoughts based on an intermittent pattern. In simpler terms, they were to have periodical pauses during test completion to reflect on and discuss the grammatical structures and accuracy features they were employing.

3.5. Procedure

3.5.1. Cognitive Model Construction. The first procedural step entailed developing a cognitive model of attributes underlying grammatical range and accuracy in IELTS. To this end, a tentative attribute list was initially developed based on the existing empirical data extracted from diagnostic studies on L2 grammatical structure (i.e., Geramipour et al., 2021; Park & Cho, 2011; Yi, 2017) and the specifications and technical reports of the grammatical range and accuracy measure provided by official developers of IELTS, including British Council (British Council website (https://www.britishcouncil.org/), IDP Education (https://ielts.idp.com/), and Cambridge Assessment English (https://www.cambridgeenglish.org/). The initially developed sketch laid the foundation for cognitive model conceptualization. The model was then enriched based on grammatical titles addressed by well-established IELTS resources, including *Collins Grammar for IELTS*, *Cambridge Grammar for IELTS*, *Essential Grammar in Use*, *Advanced Grammar in Use*, and *IELTS Band 9 Grammar Secretes*. The emerging model was expanded based on the tripartite rubric proposed by Leighton and Gierl (2007), encompassing fine-grain size (i.e., the level of precision), measurability, and instructional relevance. The emerging model included ten attributes encompassing subskills or knowledge areas specific to one or more possible class(es)of proficiency based on CEFR levels and IELTS band scores (IBS) shown in Figure 2. The model construction process proceeded with the assistance of the 19-member IELTS instructor panel. To this end, every single instructor attended the semi-structured interview and expressed their views about the specificity level, instructional relevance, measurability, and level-appropriateness of the attributes and sub-attributes in the emerging model based on their teaching experiences and the course content prescribed by institutional administrators. The instructor-elicited data helped refine the model and reach the finalized version, including two macro-level and nine micro-level attributes (see the Result section).

Figure 2
Proficiency Levels Informed the Model



3.5.2. Instrument (GRAT) Development. Following the model construction phase, GRAT was developed to address all macro- and micro-level attributes underlying the emerging model. Before item development, a tentative Q-matrix was developed based on the model's specification to inform the initial sketch of GRAT. Given the level-specific nature of the subskills underlying every attribute, the test comprised three level-appropriate booklets: Basic, Independent, and Proficient. Given the multiplicity of sub-skills underlying every single attribute and in-context nature of grammar evaluation in productive modules of IELTS, every GRAT item addressed various structural constructs underlying a single or two inter-related attribute(s) through partially lengthy texts followed by one correct and three distracting choices. The texts were truncated IELTS scripts and essays created by ChatGPT (GPT-4o). The texts ranged in potential score bands between 5 and 9 based on the booklets they belonged to. The chatbot was also consulted in modifying the surface-level errors and developing multiple choices (see sample items in the Appendix). The overall structure of each booklet accorded with the nine-attribute model; therefore, there were instances of absent or double-weighted attributes in each of the three booklets. The final version of GRAT, including 30 multiple-choice items (10 each booklet), was structured in two different formats: computer-assisted and pen-and-paper. The computer-assisted format designed by the Quize24 website (https://www.quiz24.ir/) was a time-limited (60-minute) whole-test representation format approximating the pen-and-paper version.

3.5.3. Instrument (GRAT) Verification. The five IELTS instructors who specialized in testing checked the overall structure of the GRAT draft and modified some potentially problematic texts and choices. The computer-based version of the 30-item test was administered to 17 advanced-level IELTS candidates. The analysis of item characteristics such as discrimination, facility, and readability indices indicated that four items needed removal: two from the Basic booklet, one from the Proficient booklet, and one from the Independent booklet. Additionally, one item in the Basic booklet required rewriting, and two items in the Independent booklet were merged. Consequently, the revised version of the test comprised 25 multiple-choice items distributed among the Basic (eight items), Independent (nine items), and Proficient (eight items) booklets. Based on the L2 readability index (RDL2) computed by the computational tool Coh-metrix 3.0, the average text readability decreased from the first to the third booklet (Basic: 5.89, Independent: 4.83, Proficient: 4.197), showing an ascending-order text difficulty. The IELTS candidate sample (n = 567) then took either the pen-and-paper or computer-assisted test under the direct guidance of the IELTS instructors. The test takers were allowed to answer the items in either the Basic and Proficient booklets (B1 and B2 level candidates) or all three booklets (C1 and C2 level candidates) based on their self-perceived levels of English proficiency. One hundred thirty-nine IELTS candidates out of the 567-member sample of the research perceived themselves as advanced users of English and answered the questions in all three booklets (25-item GRAT), whereas the remaining 428 only answered the items in the first two booklets (17-item GRAT).

Throughout the validation phase, the five-member instructor panel verified the attribute-item relationships in the Q-matrix based on the verbal reports from the 12 IELTS candidates who attended the think-aloud activity. The tentative Q-matrices developed for the 17- and 25-item booklets before instrument development were refined based on the expert's interpretation of the verbal reports. The response data and the item-by-attribute specifications were used to run model fit analysis through G-DINA. The response data were used to estimate the tests' internal consistency. Cronbach's α values for the three booklets (Basic: 0.76, Independent:0.69, Proficient: 0.84) showed the internal consistency of GRAT. In addition to the estimated Cronbach alpha, a couple of G-Dina parameters estimated based on the response data, including pattern accuracy (p_{a}) and consistency (p_c) of the nine attributes, corroborated the reliability of both 17-item ($p_a \geq 0.71$ and $p_c \geq .80$) and 25-item ($p_a \geq 0.78$ and $p_c \geq .83$) versions of GRAT. The internal validity of the test was also established by estimating and plotting the item-specific differences between the mastery and non-mastery proportions (see the Item Mastery Plot in the Appendix). The average of these values, representing the discriminatory power of the 17-item (0.52) and 25-item (0.60) GRAT, showed the acceptable discrimination capacity of the test (von Davier & Lee, 2019).

3.6. Data Analysis

The qualitative content analysis of interview data, specifications provided by IELTS developers and resources, and field-relevant empirical evidence yielded the cognitive model underlying GRAT (the area of inquiry in Research Question 1). The verbal think-aloud data were analyzed qualitatively to ensure the item-attribute connections addressed by GRAT based on the cognitive model developed in the introductory phase. The finalized Q-matrices and the response data were used to run a G-DINA model. The analytical modeling process was twofold: (a) estimating the model fit indices, including log-likelihood (LL), Deviance (-2LL), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Standardized Root Mean Square Residual (SRMSR), and (b) estimating the skill mastery statistics, including latent classes proportions and attribute-specific posterior probabilities. The first strand of statistics helped to ascertain whether the emerging CDA-informed instrument meets the model-data fit requirements, whereas the second one provided detailed parameters describing the test's consistency in grouping the examinees under attribute-specific mastery or non-mastery headings. The G-DINA package (Ma et al., 2016, version 2.9.4) in the RStudio software (version 2024.04.1) helped to proceed with data analysis.

4. Results

4.1. CDA Model

The first research inquiry entailed developing a cognitive model for diagnosing IELTS candidates' weaknesses and strengths in grammatical range and accuracy, a descriptive criterion for gauging IELTS writing and speaking proficiency. To this end, a multi-stage model construction procedure was followed to gather and analyze the qualitative data from the literature, official IELTS websites, well-established IELTS preparation resources, and interviews with IELTS instructors. The initial sketch of the model, grounded on literature-driven data and IELTS developers' descriptions, comprised three general attributes, including familiarity with various structural rules, being cognizant of module-specific structural nuances, and knowing how and when to use varied sentence structures (see Table 2).

Table 2
The Initial Sketch of the CDA Model

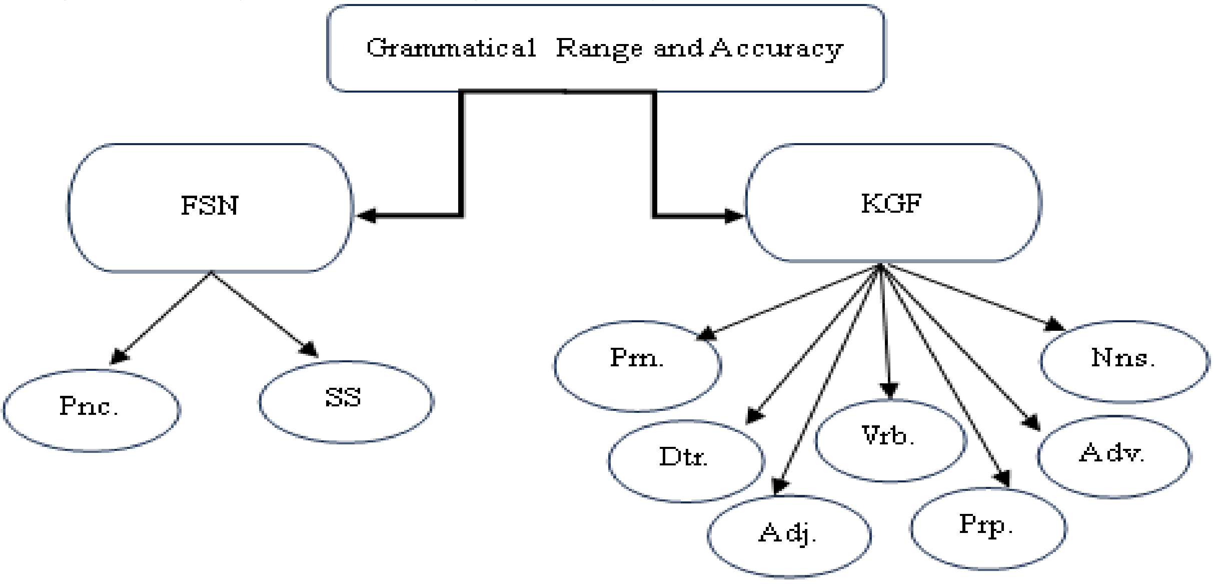
Attribute	Sub-component
Familiarity with various structural rules	• Verb Use (Tenses) • Modifiers • Prepositions • Subject-Verb Agreement
Being cognizant of module-specific (natural vs. academic language) nuances	• Idiomatic Phrases (speaking) • Active/Passive Voices (academic writing) • Punctuation (writing)
Knowing how and when to use varied sentence structure	• Simple Structure • Compound Structure • Complex Structure • Compound-Complex Structures

In the next stage, the components in the primary model were dissected based on the data collected from printed grammatical resources introduced or provided by the official IELTS web pages (see the list in the Method section). The refined model (Table 1 in the Appendix) was similar to the initial sketch in addressing general attributes but more detailed in determining the sub-skills underlying each attribute. In the subsequent step, the specifications provided by Grammar Reference on the British Council website were exploited to expand the model and reorganize the emerging attributes and subcomponents based on their scopes and proficiency thresholds. This second model refinement phase yielded a proficiency-based model, including two macro-level and ten micro-level attributes. The macro-level attributes included Knowledge of Grammatical Forms (KGF) and Familiarity with Structural Nuances (FSN). Attributes underlying KGF included knowledge of using pronouns (Prn.), determiners/quantifiers (Dtr.), possessives (Pss.), adjectives (Adj.), adverbials (Adv.), nouns/noun phrases (Nns.), verbs/tenses (Vrb.), and prepositions (Prp.). On the other hand, FSN encompassed familiarity with structural sophistication (SS) and punctuation (Pnc.) (see Table 2 in the Appendix).

The proficiency-based model was refined more based on the results drawn from the content analysis of the IELTS instructors' interviews. Throughout this concluding phase, the fine-grain size, measurability, and instructional relevance of the subcomponents underlying the model were examined based on instructor-elicited data. As the interview data revealed, though no IELTS preparation course curriculum addresses the grammatical components tailored to A1 and A2 levels, the items should not be excluded from the model, given that all IELTS candidates attending these courses are to show good mastery of these structures. A thematic analysis of the interview data called for several instances of attribute integration (nouns and noun phrases), attribute omission (possessives), level change (e.g., gerunds and infinitives), component addition (e.g., causative verbs), and component replacement (e.g., possessive pronouns). Subsequently, all the changes above (the color-directed notes beneath Table 2 in the Appendix) were made, and a nine-attribute model, as illustrated in Figure 3, emerged.

Figure 3

A Cognitive Model of Grammatical Range and Accuracy



4.2. Model Fit Parameters

Tables 3 and 4 below display the refined item-attribute relationships in the two matrices developed to analyze the response data elicited from the advanced (Q-matrix 1) and intermediate/upper intermediate (Q-matrix 2) participants.

Table 3
Q-matrix 1

Booklet	Item	Attribute								
		FSN							KGF	
		1	2	3	4	5	6	7	8	9
Basic-Level	1	1	0	0	0	0	0	0	0	0
	2	0	1	0	0	0	0	0	0	0
	3	0	0	1	0	0	0	0	0	0
	4	0	0	0	1	0	0	0	0	0
	5	0	0	0	0	1	0	0	0	0
	6	0	0	0	0	0	1	0	0	0
	7	0	0	0	0	0	1	0	0	0
	8	0	0	0	0	0	0	1	0	0
Independent-Level	9	1	0	0	0	0	0	0	0	0
	10	0	1	1	0	0	0	0	0	0
	11	0	0	0	1	0	0	0	0	0
	12	0	0	0	0	1	0	0	0	0
	13	0	0	0	0	0	1	0	0	0
	14	0	0	0	0	0	1	0	0	0
	15	0	0	0	0	0	0	1	0	0
	16	0	0	0	0	0	0	0	1	1
	17	0	0	0	0	0	0	0	1	1
Proficient-Level	18	1	0	0	0	0	0	0	0	0
	19	0	1	1	0	0	0	0	0	0
	20	0	0	0	1	0	0	0	0	0
	21	0	0	0	0	1	0	0	0	0
	22	0	0	0	0	0	1	0	0	0
	23	0	0	0	0	0	1	0	0	0
	24	0	0	0	0	0	0	1	0	0
	25	0	0	0	0	0	0	0	1	1

Table 4
Q-matrix 2

Booklet	Item	Attribute								
		FSN							KGF	
		1	2	3	4	5	6	7	8	9
Basic-Level	1	1	0	0	0	0	0	0	0	0
	2	0	1	0	0	0	0	0	0	0
	3	0	0	1	0	0	0	0	0	0
	4	0	0	0	1	0	0	0	0	0
	5	0	0	0	0	1	0	0	0	0
	6	0	0	0	0	0	1	0	0	0
	7	0	0	0	0	0	1	0	0	0
	8	0	0	0	0	0	0	1	0	0
Independent-Level	9	1	0	0	0	0	0	0	0	0
	10	0	1	1	0	0	0	0	0	0
	11	0	0	0	1	0	0	0	0	0
	12	0	0	0	0	1	0	0	0	0
	13	0	0	0	0	0	1	0	0	0
	14	0	0	0	0	0	1	0	0	0
	15	0	0	0	0	0	0	1	0	0
	16	0	0	0	0	0	0	0	1	1
	17	0	0	0	0	0	0	0	1	1

Both matrices above included all nine micro-level attributes, including pronouns, determinants/quantifiers, adjectives, adverbials, nouns/noun phrases, verbs/tenses, prepositions, structural sophistication, and punctuation, represented with numbers 1 to 9, respectively. The Q-matrices above and the corresponding response data were used to examine the model fit level based on the G-DINA model. As seen in Table 5, the values estimated for the absolute fit index (SRMSR values) were lower than 0.05, indicating that the average discrepancy between observed and predicted correlations was negligible (Maydeu-Olivares, 2013). This result suggests a good fit for the 17- and 25-item versions of GRAT. The relative fit statistics in Table 5 (LL, -2LL, AIC, and BIC) corroborated the closer model fit for the response data associated with Q-matrix 1 given that higher log-likelihood (LL) and lower deviance (-2LL) and information criteria (AIC and BIC) values predict a better fit. In sum, the absolute and relative fit indices testified to the satisfactory fit of the G-Dina to the data.

Table 5
Absolute and Relative Fit Statistics

Q-matrix	Npar	LL	-2LL	AIC	BIC	SRMSR
Q-matrix 1	571	-1322.90	2645.81	3787.81	5463.39	.0444
Q-matrix 2	551	-3396.88	6793.76	7895.76	10132.34	.0157

Table 6 shows the absolute fit indices for the 17-item and 25-item versions of GRAT in terms of different statistical measures used in a psychometric context. The statistics estimated for Proportion Correct (also called item difficulty) show an acceptable degree of model fit, as both central (mean) and dispersion (max) measures were low (close to zero), and the p-values were higher than .05. This result showed that the observed data were consistent with what the model predicted in terms of item difficulty. The max (z.stat) and p-values ($p < .05$) of Transformed Correlations and Log Odds Ratio, the measures that normalize data for statistical analysis and quantify the strength of binary variables' association, respectively, showed that before accounting for multiple comparisons, the item did not fit the model well. Nevertheless, the fact that adjusted p-values were higher than 0.05 showed no significant deviation from the model. Accordingly, the item-level fit indices testified that there was no clear evidence of items fitting poorly in the model in the two versions of GRAT.

Table 6

Absolute Item-based Fit Statistics

Version	Measure	Mean (stats)	Max (stats)	Max (z.stats)	p-value	Adj. p-value
17-item	Proportion Correct	0.001	0.003	0.197	0.843	1.000
	Transformed Corr.	0.029	0.267	2.512	0.000	0.061
	Log Odds Ratio	0.219	0.807	1.159	0.000	0.090
25-item	Proportion Correct	0.001	0.002	0.063	0.949	1.000
	Transformed Corr.	0.059	0.308	3.577	0.000	0.100
	Log Odds Ratio	0.577	3.588	2.826	0.004	1.000

4.3. Skill Mastery Parameters

Table 7 displays the most frequent latent classes of attribute mastery among all 512 (2^9) classes possible. The inclusion criterion for all the classes was a more-than-one observation percentage. As seen in the table, the thorough mastery pattern, which means the mastery of all nine attributes, was the one that included the highest proportion of the intermediate/upper intermediate (12.60%) and advanced (16.55%) examinees. The other pattern common among those who took the 17-item booklet was the two-attribute non-mastery pattern (10.22%), showing the difficulty of 10.22% of the examinees only in learning the two attributes underlying FSN. The other noteworthy attribute mastery classes included a couple of three-attribute non-mastery (17-item: 111110100 and 101111100, 25-item:110101110 and 011111100, 25) and one-attribute non-mastery (17-item: 111111101 and 1111111110, 25-item: 110101110 and 011111100) patterns. One or both of the attributes underlying FSN corresponded to a non-mastery state (0) in these classes. According to Table 7, the patterns with more than four non-mastered attributes were the least observed classes, specifically among the advanced examiners (less than 20%).

Table 7

Posterior Probability Percentages for Frequent AMPs

Pattern	17-item GRAT			25-item GRAT		
	Rank	Latent Class	%	Rank	Latent Class	%
Full mastery	1	111111111	12.60	1	111111111	16.55
One-attribute non-mastery	5	111111101	2.49	4	110111111	4.05
	5	111111110	2.49	5	111110111	3.53
	6	111111011	2.34	6	111111011	2.25
	8	111110111	2.19	8	101111111	1.65
	---	---	---	12	111011111	1.41
	---	---	---	13	011111111	1.31
Two-attribute non-mastery	2	111111100	10.20	4	110111101	4.05
	10	111010111	1.87	5	111110110	3.53
	12	011011111	1.54	5	111110101	3.53
	16	111110101	1.07	6	111111001	2.25
	16	110110110	1.07	6	111111010	2.25
	---	---	---	8	101111110	1.65
	---	---	---	8	101111101	1.65
	---	---	---	11	110111011	1.43
	---	---	---	14	011111101	1.31
	---	---	---	14	011111110	1.31

	---	---	---	15	001111111	1.22
Three-attribute non-mastery	3	111110100	3.37	2	110101110	5.77
	4	101111100	3.36	3	011111100	4.33
	7	110011100	2.25	7	011110101	2.22
	9	011111100	2.08	7	011110110	2.22
	11	111101100	1.63	9	011110011	1.63
	11	111111000	1.63	11	110111001	1.43
	12	110111000	1.54	11	110111010	1.43
	---	---	---	14	001111110	1.22
	---	---	---	14	001111101	1.22
	---	---	---	17	111011100	1.12
Four-attribute non-mastery	13	111010100	1.17	10	110110100	1.46
	14	111011000	1.16	13	011011000	1.32
	15	101110100	1.11	16	110111000	1.17
		Total	57.16		Total	81.47

The probability coefficients in Table 8 show the likelihood that the examinees who took the 17- and 25-item booklets achieved mastery in each of the nine micro-level attributes in the CDA model. As suggested by the results, among the attributes underlying knowledge of grammatical forms, nouns/noun phrases were the attributes mastered by the vast majority of the intermediate/upper intermediate (80.19%) and advanced (92.09%) examinees. Knowledge of pronoun use was another skill highly mastered by the intermediate/upper-intermediate (75.99%) and advanced examinees (85.65%). The mastery of all other attributes underlying KGF was found to be of moderate probability. Knowledge of verbs/tenses use, mastered by 65.24% of the intermediate/upper-intermediate and 63.83% of the advanced examinees, and the two attributes underlying familiarity with structural nuances, learned by 59.94% of intermediate/upper-intermediate and 61.80% of the advanced examinees, enjoyed the lowest mastery probabilities.

Table 8
Skill Mastery Probabilities

Macro-level Attribute	Micro-level Attribute	Probability	
		17-item	25-item
KGF	Pro.	.7599	.8566
	Det.	.6734	.7681
	Adj.	.7820	.6696
	Adv.	.6921	.8065
	Non.	.8019	9209
	Vrb.	.6524	6383
	Prp.	.7365	6834
FSN	Str.	.5994	6180
	Pun.	.5994	6180

5. Discussion

The primary objective of this study was to construct a model delineating cognitive attributes fundamental to an IELTS assessment construct, focusing on grammatical range and accuracy. This model aimed to lay the groundwork for developing a CDA-informed instrument named GRAT, designed to discern candidates' grammatical strengths and weaknesses critical for achieving success in the IELTS examination. Employing a multi-stage qualitative methodology involving data collection, interpretation, and synthesis, a dual-tiered framework emerged. This framework categorized cognitive attributes underlying the testing construct into two overarching domains: (a) comprehension of grammatical forms, encompassing pronouns, determiners/quantifiers, adjectives, adverbials, nouns/noun phrases, verbs/tenses, and prepositions, and (b) familiarity with structural nuances, including punctuation and structural sophistication.

Each micro-level attribute incorporated several level-specific sub-components tailored to various CEFR proficiency classes, including A1-A2, B1-B2, and C1-C2.

Despite the innovative approach (genuinely diagnostic) and the novel scope (grammatical range and accuracy in IELTS), which posed challenges in establishing the model's credibility against existing empirical evidence, it was noted that the emerging model shared common micro-level attributes (e.g., use of verbs and prepositions) with infrequently encountered retrofitted-design studies on cognitive diagnostic assessments of grammar in high-stakes English proficiency tests (Geramipour et al., 2021; Park & Cho, 2011). Nonetheless, the domain-specific model derived from the multi-stage inductive process of this study demonstrated superior comprehensiveness not only in addressing intellectual skills relevant to diverse grammatical forms but also in encompassing cognitive resources necessary for generating precise and sophisticated language constructs at superficial (punctuation) and deep (structural complexity) levels.

The intricate and nuanced structure of the model precisely delineated the content and framework of the CDA-informed instrument. Nevertheless, the Q-matrices and initial test drafts underwent continuous refinement throughout the iterative process of model development. The specific design of the test, wherein items predominantly explored distinct sub-components representing a single attribute or closely related attributes within a macro-level, was validated through examinee verbal reports, expert evaluations, and model fit parameters computed based on G-DINA assumptions. A crucial consideration lies in enhancing the credibility of the testing framework by incorporating more frequent items, each targeting specific sub-components outlined in the model. Test administration would be supported by computer algorithms generating different test versions composed of randomly selected items drawn from a comprehensive item pool addressing various sub-skills within the model to mitigate potential respondent fatigue and disinterest stemming from an excessive number of test items. Such computer-assisted randomized formats of GRAT could also counter the potential for repeated exposure effects associated with using the same test format.

The study also explored GRAT's ability to generate reliable and valid data for diagnosing language mastery. The full 25-item version demonstrated superior performance compared to a shorter version, with G-DINA analysis showing good model fit for both. Notably, the full format facilitated more precise classification, placing over 80% of examinees into distinct mastery profiles. Item analysis revealed strong discrimination between mastery and non-mastery states, even for items with lower discriminatory power. These findings, supported by von Davier and Lee (2019), suggest GRAT's potential as a diagnostic tool. However, the limited sample size necessitates further research with larger cohorts to confirm item efficacy and optimize the test.

Similar to a substantial body of genuinely diagnostic (e.g., Li et al., 2021; Ma & Meng, 2014; Toprak & Cakir, 2021; Wei, 2016) and retrofitted design (Effatpanah, 2019; Mohammed et al., 2023) studies in diagnostic L2 assessment, the findings of this study revealed that the full mastery pattern, which indicates proficiency in all cognitive attributes assessed by the test, was either the sole or one of the most frequent patterns observed. Additionally, a notable proportion of intermediate/upper-intermediate or advanced examinees fell into patterns indicating non-mastery of one or both attributes related to familiarity with structural nuances. This outcome, coupled with the lowest mastery probabilities observed for these attributes, underscores GRAT's capacity to highlight structural weaknesses among examinees. Another significant finding was the prevalent non-mastery of knowledge related to various verbs/tenses across examinees of different proficiency levels. Conversely, proficiency in using nouns/noun phrases and pronouns emerged as the top two attributes mastered by examinees, regardless of their proficiency levels.

The findings regarding attribute mastery align closely with the expectations of IELTS authorities and instructors expressed throughout the research process. This alignment underscores GRAT's potential diagnostic value, echoing Roussos et al.'s (2007) assertion that the effectiveness of mastery/non-mastery models hinges on their alignment with user expectations. The congruence observed may be attributed to the inherent complexity and acquisition challenges associated with clause- and sentence-level structures such as tenses and inversion, contrasting with the more manageable and straightforward word- and phrase-level elements. Furthermore, the sequence of attribute mastery also reflects stages of grammatical development as outlined in Pienemann's

Processability Theory (1998), where procedures involving verbs/phrases and subordinate clauses prove more cognitively demanding compared to more straightforward procedures like lexical lemma recognition and noun/phrase construction. This interpretation is supported by local empirical data presented by Dehghani and Bagheri (2016), highlighting the complexity of mastering grammatical sub-components related to verb/tense usage (e.g., conditional and passive structures) and structural nuances (e.g., reported speech) among Iranian EFL learners.

This study reinforced the suitability of G-DINA, a saturated model, for analyzing diagnostic L2 grammar tests. The satisfactory model fit and consistent classification accuracy demonstrated its effectiveness in capturing the interplay between various grammar sub-skills. This finding aligns with prior research in L2 reading (Toprak & Cakir, 2021), listening (Ma & Meng, 2014), and writing (Effatpanah et al., 2019), where G-DINA successfully modeled the complex relationships among skills. Notably, the model's flexibility in accommodating both compensatory and non-compensatory interactions mirrors the multifaceted nature of L2 grammar acquisition (Chen & Chen, 2016; Ravand, 2016). This further strengthens its position as a preferred choice when the optimal model remains unknown (Chen et al., 2013; Li et al., 2016) and complements existing evidence for its application in high-stakes grammar assessments (Geramipour et al., 2021). However, the non-comparative design of this study, coupled with support for compensatory reparametrized unified models (C-RUM) in analyzing grammar tests (Yi, 2017), necessitates further research to definitively identify the optimal modeling approach.

6. Conclusion

The current mixed-methods exploratory research elucidated the intricate cognitive framework underlying grammatical range and accuracy in IELTS through a genuinely diagnostic approach to cognitive diagnostic assessment. The emergent cognitive model, featuring two macro-level attributes—Knowledge of Grammatical Forms and Familiarity with Structural Nuances —and nine micro-level attributes, offered a detailed comprehension of the competencies crucial for mastering the grammar base required to thrive in IELTS. The development, piloting, and validation of the CDA-informed test, GRAT, using the G-DINA model, underscored the innovative measure's robustness and diagnostic precision. The model's satisfactory fit and the robust skill mastery parameters affirmed GRAT's efficacy in accurately categorizing examinees based on their attribute-specific mastery levels. Significantly, the less mastered attributes, such as punctuation, structural sophistication, and verb tense usage, provided critical insights for enhancing IELTS preparation courses. This study not only contributed to the L2 assessment field but also set a new standard for diagnostic precision in evaluating grammatical competencies, promising a transformative impact on both teaching and testing practices. This work not only contributes to the field of L2 assessment but also sets a new standard for diagnostic grammar evaluation, potentially impacting both teaching and testing practices. The identified cognitive model offers a targeted framework for educators, while the study highlights the need for a more focused approach to specific areas within IELTS preparation.

While the findings of this study are promising, it is vital to acknowledge certain limitations. The sample size, though substantial, may not fully represent the diverse population of IELTS test-takers. Additionally, the pre-fabricated test format may limit the flexibility and adaptability of the assessment. Future research should consider larger, more diverse samples and explore the development of a flexible, item-pool-based assessment to enhance diagnostic accuracy. Furthermore, longitudinal studies investigating the long-term impact of targeted instruction based on the identified cognitive attributes could provide valuable insights into the practical applications of this research. Finally, the dual administration of the test in paper-and-pencil and computer-based formats raises concerns about mode effects. Although the study assumes equivalence, research indicates that different modes can introduce construct-irrelevant factors. Future research should examine this equivalence to reduce potential biases.

The implications of this study are profound. The refined cognitive model may offer educators a targeted framework to address specific grammatical weaknesses in students, thereby optimizing instructional strategies and improving learning outcomes. Pedagogically, the findings suggest the need for a more focused approach to teaching punctuation, structural sophistication, and verb tenses within IELTS preparation courses. Nonetheless, the study is not without limitations. The limited sample size (n = 567) may not capture the full diversity of IELTS test-takers. In addition, using a pre-fabricated test including items with fragmentary choices addressing several sub-skills underlying one or more attributes, instead of a dynamic, programmed test with single-construct items selected randomly from an item pool tailored to specific sub-skills, may limit the assessment's adaptability and precision. Future research should consider more diverse and representative respondent samples and the development of a more flexible, item-pool-based assessment system to enhance diagnostic accuracy. Moreover, longitudinal studies could explore the long-term impact of targeted grammatical instruction based on the identified cognitive attributes, thereby solidifying the practical applications of this research.

Acknowledgments

The authors express their gratitude to all of the individuals who either participated in the study or contributed to improve the quality of the manuscript.

Declaration of Conflicting Interests (obligatory)

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding (obligatory)

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of AI-Generated Content

The authors made limited use of AI-generated technologies, employing them solely to enhance readability and language quality while maintaining human oversight and control to prevent errors, omissions, or biased content.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. New York: Continuum.

Choi, H-J., Rupp, A. A., & Pan, M. (2012). Standardized diagnostic assessment design and analysis: Key ideas from modern measurement theory. In M. C. Mok (ed.), *Self directed learning oriented assessments in the Asia Pacific* (pp.61–85). The Education University of Hong Kong. https://doi.org/ 10.1007/978-94-007-4507-0_4

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). *Thousand Oaks*, SAGE.

de Bot, K., Lowie, W., & Verspoor, M. (2007). A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and Cognition*, *10*(1), 7–21. https://doi.org/10.1017/S1366728906002732

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199. ttps://doi.org/10.1007/s11336-011-9207-7

Dehghani, A. P., & Bagheri, M. B. (2016). Investigating difficulty order of certain English grammar features in an Iranian EFL setting. *International Journal of English Linguistics*, *6*(6), 209–220. http://doi.org/10.5539/ijel.v6n6p209

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). 31A review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (pp. 979–1030), Elsevier.

Effatpanah, F. (2019). Application of cognitive diagnostic models to the listening section of the International English Language Testing System (IELTS). *International Journal of Language Testing*, *9*(1), 1–28.

Effatpanah, F., Baghaei, P., & Boori, A. A. (2019). Diagnosing EFL learners' writing ability: A diagnostic classification modeling analysis. *Language Testing in Asia*, *9*(12), 1–23. https://doi.org/10.1186/s40468-019-0090-y

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380–396. https://doi.org/10.1037/1082-989X.3.3.380

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*(4), 343–368. https://doi.org/10.1111/j.1745-3984.2001.tb01131.x

Ellis, R. (2006). Current issues in the teaching of grammar: An SLA perspective. In R. Ellis (Ed.), *Language acquisition and language socialization: Ecological perspectives* (pp. 88–107). Routledge.

Ellis, N. C. (2016). Language acquisition as a complex adaptive system. In J. P. Lantolf & M. E. Poehner (Eds.), *The Routledge handbook of sociocultural theory and second language development* (pp. 34–46). Routledge.

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/s40468-02000117-y

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

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral thesis). University of Illinois at Urbana-Champaign.

Jang, E. E. (2008). A Review of cognitive diagnostic assessment for education: Theory and application. *International Journal of Testing*, *8*(3), 290–295. https://doi.org/10.1080/15305050802262332

Jang, E. E. (2009). 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/10.1177/0265532208097374

Kargar Behbahani, H., Darazi, M. A., & Kumari R, L. (2024). Developing Iranian EFL learners' grammatical knowledge: Insights from spaced versus massed instruction. *Journal of Languages and Language Teaching*, *12*(2), 612. https://doi.org/10.33394/jollt.v12i2.10296

Ketabi, S., Alavi, S. M., & Ravand, H. Diagnostic test construction: insights from cognitive diagnostic modeling. *International Journal of Language Testing*, *11*(1), 22–35.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. Longman.

Larsen-Freeman, D., & Celce-Murcia, M. (2015). *The grammar book: Form, meaning, and use for English language teachers* (3rd ed.). National Geographic Learning/Cengage Learning.

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., & Hunka, S. M. (2004). The attribute hierarchy model for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*, 205–237. https://doi.org/10.1111/j.1745-3984.2004.tb01163.x

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511611186

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. Language Teaching Publications.

Li, J., Mao, X., & Zhang, X. (2021). Q-matrix estimation (validation) methods for cognitive diagnosis. *Advances in Psychological Science*, *29*(12), 2272–2280. https://doi.org/10.3724/SP.J.1042.2021.02272

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press.

Lord, F. M. (1952). *A theory of test scores*. Psychometric Society.

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). G-DINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *74*(10), 1–26. https://doi.org/10.18637/jss.v074.i10

Ma, X., & Meng, Y. (2014). Towards personalized English learning diagnosis: Cognitive diagnostic modelling for EFL listening. *Asian Journal of Education and e-Learning*, *2*(5), 336–348.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models (with discussion). Measurement: *Interdisciplinary Research and Perspectives*, *11*, 71–137. https://doi.org/10.1080/15366367.2013.831680

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*(4), 439–483. https://doi.org/10.1007/BF02294388

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*(4), 477–496. https://doi.org/10.1191/0265532202lt241oa

Mohammed, A., Shareef Dawood, A. K., Alghazali, T., Kadhim, Q. K., Sabti, 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*, 1, 1–20. https://doi.org/10.22034/IJLT.2022.362849.1195

Park, C., & Cho, S. (2011). Cognitive diagnostic writing assessment for Korean learners of English. *English Teaching 66*(4), 101–117. https://doi.org/10.15858/engtea.66.4.201112.101

Pilcher, N, & Richards, K. (2017). Challenging the power invested in the international English language testing system (IELTS): Why determining 'English' preparedness needs to be undertaken within the subject context. *Power and Education*, *9*(1), 3–7. https://doi.org/10.1177/1757743817691995

Pishghadam, R., & Miri, M. A. (2021). Toward an emotioncy-based education: Asystematic review of the literature. *Frontiers in Education*, 6, 606–619. https://doi.org/10.3389/feduc.2021.606619

Rahimi, M., & Zhang, L. (2015). Exploring non-native English-speaking teachers' cognitions about corrective feedback in teaching English oral communication. *System*, *55*, 111–122. https://doi.org/10.1016/j.system.2015.09.006

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

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-theart. *Measurement: Interdisciplinary Research and Perspectives*, *6*(4), 219–262. https://doi.org/10.1080/15366360802490866

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

Ruzmetova, D. K. (2024). Crucial role of grammar in effective communication. *Mental Enlightenment Scientific-Methodological Journal*, *5*(05), 246–250. https://doi.org/10.37547/mesmj-V5-I5-34%20

Shintani, N., & Ellis, R. (2018). The comparative effect of metalinguisticexplanation and direct written corrective feedback on learners' explicit and implicit knowledge of the English indefinite article. *Journal of Second Language Writing*, *22*(3), 286–306. https://doi.org/10.1016/j.jslw.2013.04.002

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101. https://doi.org/10.2307/1422689

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*, 106–131. https://doi.org/10.1177/0265532220941470

von Davier, M. (2005). *A general diagnostic model applied to language testing data*. Educational Testing Service. https://doi.org/10.1348/000711007X193957

von Davier, M., & Lee, Y. S. (2019). *Handbook of diagnostic classification models: Models and model extensions, applications, software packages*. Springer

Wei, J. (2016). *Bridging the gap between research and practice: construction and validation of a CDA-informed English reading test for China's twelfth graders* (Unpublished Doctoral Dissertation). University of Illinois at Urbana-Champaign.

Williamson, J. (2023). *Cognitive diagnostic models and how they can be useful*. Cambridge.

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

Zhang, S., & Thompson, N. (2004). A diagnostic language assessment system (review). *Canadian Modern Language Review*, *61*(2), 290–293. https://doi.org/10.1353/cml.2005.0011

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*, *16*(29), 1–7. https://doi.org/10.1088/1742-6596/1629/1/012037

Appendix
Sample Items of GRAT

Basic Booklet

1. Read the text and select the option that fits correctly in the blank spaces. The symbol ( ) indicates where no word is needed.

The weather wasn't so nice yesterday. … was raining and … was a strong wind. I knew it would be deadly dull to spend the whole day alone by …, so I called a bunch of … close friends and asked them to attend a small party in my garden. I told them to ask some of … friends that could cheer … As I know, all my friends like such unplanned parties, except for the surprise … in the past summer I held for my husband's birthday.

a) it, there, myself, my, their, us, one
b) that, it, me, mine, ( ), them, ( )
c) that, it, myself, mine, their, us, one
d) it, there, me, my, ( ), them, ( )

Proficient Booklet

1. Read the text and select the option that fits correctly in the blank spaces.

After Tom's accident, both his parents felt guilty but blamed … They criticized … conduct while hanging out in the police station. After a heated argument, the father called Peter, a close friend of … and asked about the last time he had seen Tom. Peter told them that Tom had left him around 2:30 (p.m.) to walk home by …

a) each other, their, his, himself
b) one another, his, Tom, him
c) one another, each other's, Tom's, himself
d) themselves, each other's, his, his

Independent Booklet

1. Read the following email carefully. The multiple-choice items beneath the text confirms or provides alternatives for italicized word/words. Decide on the correct item.

Hi Dana. Yes, Jan's a lot better, thanks. We (1) got vaccinated ourselves against hepatitis before we went to West Africa, so Jan was just unlucky to get it. He went into work after we go back although he was feeling bad, and some of his colleagues were worried about (2) *getting it themselves*. By coincidence, his boss said that (3) *he'd caught himself hepatitis* a few years ago. When he's completely recovered, (4) *John and myself* are left to Paris for a few days -if I can get Jan (5) *to tear him away* from his office!
Must go now. The children have just shouted that they wand some juice and (6) *they can't reach it themselves*.

a) got vaccinated/getting it by themselves/ ✓/ John and I/ to tear away himself /they can't reach it by themselves
b) got ourselves vaccinated/✓/he'd caught hepatitis himself/ ✓/to tear himself away/✓
c) got ourselves vaccinated / getting it themselves / ✓ / John and me / to tear himself away / they can't reach it by themselves
d) b) got vaccinated / getting it by their own / ✓ / John and I / to tear away himself / ✓

Table 1
Model Refined based on the data from Printed IELTS Instructional Resources

Attribute	Sub-component	Example
Familiarity with various structural rules	• Verbs/Tenses	• Simple, continuous, and perfect aspects of the present, past, and future time
	• Agreement	• Countable and uncountable nouns, this/that and these/those, both/either/neither, each/every/all, object vs. subject NP
	• Nouns/Noun Phrases (NPs)	• Countable, uncountable, exceptions, quantifiers
	• Pronouns	• Personal, possessive, time/place, such, this/that and these/those
	• Articles	• Definite, indefinite, no article
	• Modifiers	• Adjectives, adverbs, demonstratives, possessive determiners, prepositional phrases, degree modifiers, and intensifiers
	• Comparison Rules	• Comparatives, superlatives, equal/unequal comparison modifiers
	• Prepositions	• Simple, double, compound, participle, and phrase prepositions
	• Modals	• Modal forms for the present and future (normal/strong obligation, ability/inability, possibility/impossibility, and recommendation), and past (deduction, possibility, ability/inability, strong objective obligation, matters of choice, recommendation, or regret)
	• Conditionals	• Zero (factual/true information), 1st (to imagine a condition-specific future situation), 2nd (to imagine impossible a present/future situation), 3rd (to imagine a different past situation), and (hypothetical/unreal situation in the present/future that are connected to a hypothetical/unreal situation in the past) mixed conditionals
	• Reflexive Clauses	• Defining vs Non-defining and Subject vs. Object RCs
Knowing how and when to use varied sentence structure	• Word Order	• Principles for declarative, imperative, negative, interrogative, and inverted sentences
	• Text/paragraph structure	• Main and supporting sentences in different IELTS writing genres (Agree/Disagree, Advantages/Disadvantages, Problem/Solution, Discussion, and Two-part Question)
	• Reported Speech	• Tense, pronoun, and time word change in reported statements and questions
Being cognizant of module-specific (natural vs. academic language) nuances	• Punctuation	• Commas, Colons, Semi Colons, apostrophes, Ellipsis, etc.
	• Passive Forms	• Active vs. passive structures • The Active → Passive transition • Passive forms with ‘Get’

Table 2

Model Refined based on the Data from Online IELTS Instructional Resources

Macro-Level Attribute	Micro-Level Attribute	Sub-components	CEFR Level
Knowledge of grammatical Forms	A1: Pronouns	• Personal pronouns • Indefinite pronouns • ‘It’ and ‘there’ as dummy subjects • Demonstratives • ‘One’ and ‘ones’ • Pronouns in question	A1-A2
		• Reciprocal pronouns • Reciprocal pronouns + ’s	B1-B2
		• Relative pronouns • Possessive pronouns	A1-A2 B1-B2
		• Reflexive pronouns	A1-A2 B1-B2 C1-C2
	A2: Determiners and Quantifiers	• Specific determiners • General determiners • No determiner • The indefinite article	A1-A2
		• Interrogative Determiners	B1-B2
		• Quantifiers • The definite article	A1-A2 B1-B2
	A3: Possessives	• Nouns + ’s/’ • Possessive adjectives • Questions (whose)	A1-A2
		• Reciprocal pronouns + ’s	B1-B2
		• Possessive pronouns	A1-A2 B1-B2
	A4: Adjectives	• The position of adjectives • ‘ing’ adjectives vs. ‘ed’ adjectives • Comparative and superlative adjectives • Possessive a1`djectives • Noun modifiers	A1-A2
		• Adjective order • Using adjectives as nouns • Specific position adjectives • Intensifiers and mitigators	A1-A2 B1-B2 C1-C2
	A5: Adverbials	• Typology and position • Adverbials of time • Adverbials of probability	A1-A2
		• Adverbials of manner • Adverbials of place (direction, location, distance) • Comparative adverbs	A1-A2 B1-B2
		• Intensifiers and mitigators • Superlative adverbs	B1-B2
		• Inversion after negative adverbials	C1-C2
	A6: Nouns and Noun Phrases (NPs)	• Countable and uncountable nouns • Nouns + ’s/’	A1-A2
		• Group nouns • Nouns referring to two-part things • One-part NPs (noun/pronoun) • Pre-modifiers and post-determiners	B1-B2
		• Compound nouns and possessive forms	C1-C2
		• Proper Nouns • Countable/uncountable nouns not following regular rules	A1-A2 B1-B2

Table 2 (Continued)

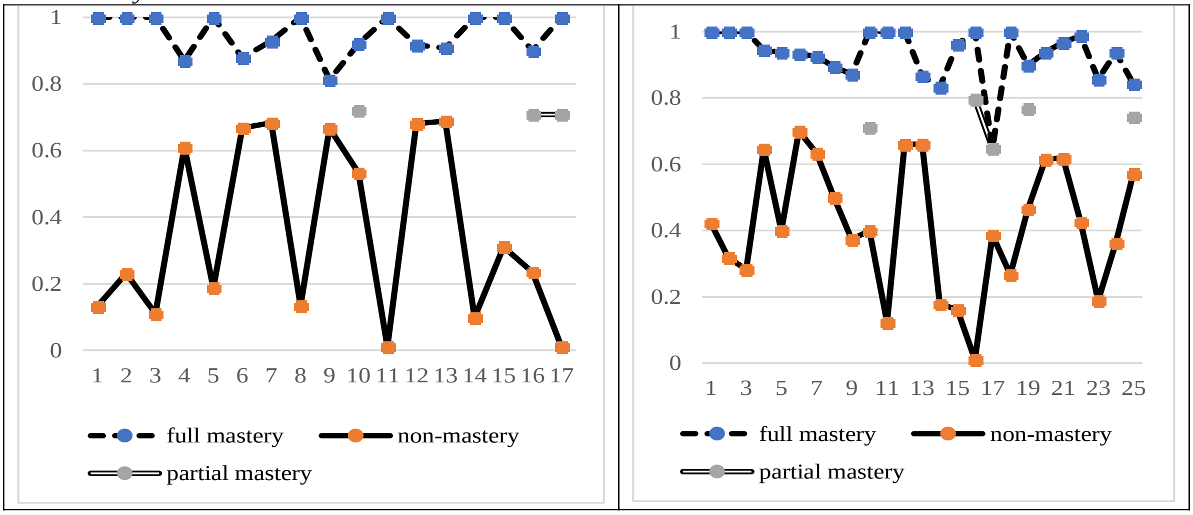
Model Refined based on the Data from Online IELTS Instructional Resources

Macro-Level Attribute	Micro-Level Attribute	Sub-components	CEFR Level
Knowledge of grammatical Forms	A7: Verbs and Tenses	• Regular/Irregular verbs • Delexical verbs • Short forms • The verb be (all forms) • Present (simple, continuous, perfect) • Past (simple and continuous) • Future (simple)	A1-A2
		• Present tense (perfect continuous) • Past tense (perfect and perfect continuous) • Future tenses and forms (continuous, perfect, be going to) • Wishes and hypothesis • *Different uses of 'used to'* • *Separable and non-separable phrasal verbs*	B1-B2
		• Specific uses of present simple and continues • Patterns with reporting verb • Causative verbs	C1-C2
		• Question and negative forms	A1-A2
		• Clause structure and verb patterns	B1-B2
		• Verb phrases • Modals • Active and passive voice • Unreal time • Verbs of senses • Gerunds and infinitives	A1-A2 B1-B2 C1-C2
		• Conditionals	B1-B2 C1-C2
	A8: Prepositions and Conjunctions	• Prepositions of time • Prepositions of place • Prepositions after adjectives • Prepositions in adverbials	A1-A2
		• Prepositions after verbs • Prepositions in relative clauses	B1-B2
		• Conjunctions in contrasting ideas • Conjunctions in participle clauses	C1-C2
Familiarity with Structural nuances	A9: Structural Sophistication	• Reported Speech	B1-B2
		• Ellipsis and substitutions • Inverted structures (negative adverbials, conditionals, and participial clauses) • Complex-compound sentences • Cleft sentences	C1-C2
	A10: Punctuation	• Rules for commas, colons, semicolons, dashes, etc.	B1-B2 C1-C2

- Level Change
- Component Addition
- Attribute Integration
- Attribute Omission
- Component replacement (from omitted attributes)

Figure 1

Item Mastery Plots


full mastery
non-mastery
partial mastery
full mastery
non-mastery
partial mastery