

Retrofitting Non-diagnostic Grammar Assessment: Application of the G-DINA Model to a High-Stakes Grammar Test

Alizadeh Birjandi, Marzieh¹, Alavi, Sayyed Mohammad^{2*}, Geramipour, Masoud³, Nayernia, Akram⁴

ARTICLE INFO

Article History:

Received: October 2025

Accepted: December 2025

KEYWORDS

Cognitive Diagnostic Model (CDM), Generalized Deterministic Inputs Noisy “And” gate (G-DINA), grammar assessment, Q-matrix, University of Tehran English Proficiency Test (UTEPT)

ABSTRACT

The purpose of this study was to compare the functioning of five Cognitive Diagnostic Models (CDMs) to identify the best-fitting CDM that can better explain the interaction underlying the attributes of the grammar section of the University of Tehran English Proficiency Test (UTEPT). To this end, a Q-matrix representing the key cognitive abilities required for the grammar section was developed. Expert input identified six essential grammar attributes: Agreements, Clauses, Lexical Knowledge, Connectors, Tense Recognition, and Voice Awareness. Data from 810 examinees (268 males and 542 females) aged 24 to 48, including those who took the test in 2022, were analyzed. The five CDMs were initially compared in terms of relative and absolute fit statistics at the test and item level to choose the best model. It was found that the Generalized Deterministic Inputs Noisy “And” gate (G-DINA) model outperformed the restrictive models; thus, it was chosen for the second phase of the study. Regarding the second purpose of the study, the G-DINA model was used to identify the strengths and weaknesses of the test takers. The results revealed that understanding Tense Recognition was the most challenging attribute for the examinees, while Lexical Knowledge was the easiest attribute. These findings highlight the need for instructional strategies focusing on enhancing Tense Recognition and integrating teaching approaches that address the interdependent nature of grammar attributes.

1. Introduction

In recent years, educational assessment has been increasingly moving towards learning the detailed cognitive profiles of students to make instruction more personalized and efficient (Rupp & Templin, 2008). Traditional testing strategies mostly provide total scores, which tend to obscure what particular skills or knowledge items the students have mastered or need to improve. This limitation has demanded the use of more sophisticated procedures of assessment to uncover the underlying learning patterns and skill mastery of each student (Embretson & Gorin, 2001).

Cognitive Diagnostic Assessment (CDA) is a comprehensive framework that has addressed this challenge by making cognitive psychology and educational measurement principles work together

¹ Faculty of Foreign Languages and Literatures, University of Tehran. Email: marzieh.alizadeh@ut.ac.ir

² Professor of Applied Linguistics, Faculty of Foreign Languages and Literatures, University of Tehran. Email: smalavi@ut.ac.ir

³ Associate Professor of Educational Data Science, Faculty of Psychology and Education, Kharazmi University, Tehran. Email: mgramipour@khu.ac.ir

⁴ English Department, Faculty of Foreign Languages, University of Science and Technology. Email: a_nayernia@iust.ac.ir

(Leighton & Gierl, 2007). Its primary goal is to identify the underlying knowledge components and cognitive abilities influencing test performance. By decomposing general abilities into distinct, measurable sub-skills or attributes, CDA enables educators to gain valuable insights into the specific areas where learners struggle, leading to more targeted support (Yi, 2017). Such detailed diagnostic feedback is crucial for designing specific teaching strategies and improving individualized learning paths (Leighton & Gierl, 2007).

In spite of the advancements in CDA, there remains a gap in the implementation of Cognitive Diagnostic Models (CDMs) in language testing. CDMs are statistical models interpreting assessment data in terms of specific skills or attributes mastered by learners (de la Torre, 2011; Sijtsma & Junker, 2006). Through the conversion of test responses into detailed profiles of skill mastery, CDMs facilitate the use of personalized instruction, adaptive testing, and more accurate diagnostic feedback compared to traditional scoring.

Despite the advancements in CDA, the application of CDMs within language assessment has mainly focused on reading comprehension tests (e.g., Boori et al., 2024; Chen & Chen, 2016; Hemmat et al., 2016; Jang, 2009; Javidanmehr & Anani, 2019; Li et al., 2015; Mohammed et al., 2023; Ravand & Robitzsch, 2018). Their use in listening comprehension has been less frequent (e.g., Aryadoust, 2018; Harding et al., 2015; Lee & Sawaki, 2009), and applications in writing assessment remain sporadic (Effatpanah et al., 2019; Kim, 2011; Ravand et al., 2024; Xie & Lei, 2021). Notably, the application of CDMs to grammar assessment is quite rare (as noted by Barghi et al., 2025; Shafipour et al., 2021; Yi, 2017). This is an important gap because grammar is a key part of language proficiency, crucial for effective communication and academic success (Ellis, 2006; VanPatten, 2017).

Selecting the most appropriate CDM is a persistent challenge in diagnostic assessment research (Jiao, 2009). Early studies often used a single CDM, usually across all test items, without addressing the model's appropriateness for the specific assessment setting. However, more recent research has shifted toward empirically comparing several CDMs in order to find the best-fitting models suited for certain data sets and testing situations. Effective model selection still requires systematic research in various kinds of evaluation scenarios (Shafipour et al., 2021).

Therefore, this study aims to advance the application of diagnostic modeling in grammar assessment by empirically comparing five CDMs to find the optimal model: the Generalized Deterministic Inputs Noisy "And" gate (G-DINA) model, the Deterministic Inputs Noisy "And" gate (DINA) model, the Deterministic Inputs Noisy "Or" gate (DINO) model, the Additive Cognitive Diagnostic Model (A-CDM), and the Reduced Reparametrized Unified Model (R-RUM), as well as validating a Q-matrix framework for high-stakes testing.

2. Review of the Literature

2.1. Applications of CDMs in Language Assessment

CDMs play a crucial role in identifying which language skills learners find easy or difficult, offering valuable feedback to inform both teaching and assessment. For instance, Moghaddam et al. (2016) examined 250 intermediate learners using a 28-item grammar test, reporting notable challenges with morphology, conjugation, syntax, and coherence rules, although learners demonstrated partial mastery of vocabulary-related skills. Similarly, Yi (2017) investigated 500 English learners through a 40-item grammar test targeting various cognitive attributes and found that morpho-syntactic knowledge—especially inflectional morphology and syntax—was the most difficult area, while vocabulary skills were relatively easier to acquire. In a study involving 300 graduate-level examinees, Geramipour et al. (2021) identified verb tense recognition as a persistent area of difficulty, with idiomatic expressions being relatively easier. Shafipour et al. (2021) analyzed data from 1,773 university students across diverse majors who completed a purpose-built General English Achievement Test. Their findings corroborated that morpho-syntactic proficiency poses greater challenges than lexical or vocabulary-related skills.

To provide the necessary theoretical foundation, it is essential to acknowledge cognitive theories of grammar acquisition, which indicate that grammar competence arises from a complex, hierarchical, and interactive system of cognitive abilities rather than isolated skills (e.g., Ellis, 2006; VanPatten, 2017). Grammar ability is best conceptualized as a multidimensional construct composed of distinct latent attributes. Cognitive diagnostic theory supports this view by emphasizing the need to

identify mastery of these specific sub-skills to provide diagnostic feedback that informs teaching and learning (Yi, 2017).

As previously noted, CDMs help teachers and testers understand exactly which skills a learner has mastered and which need improvement. Among these models, the G-DINA as a saturated model is popular because it is flexible—it can handle different ways attributes interact, whether they work together or separately, which makes it ideal for analyzing complex learning skills and components like reading and grammar. This flexibility helps provide detailed, accurate feedback that can guide better teaching and learning (Mei & Chen, 2024).

Supporting these points, Mirzaei et al. (2020) retrofitted the IELTS reading test with the G-DINA model, identifying six key reading attributes and revealing significant variation in mastery profiles across 1025 test-takers. Similarly, Shafipoor et al. (2021) compared six CDMs—including G-DINA, LLM, R-RUM, ACDM, DINA, and DINO—using grammar and vocabulary data from a dedicated English achievement test. Their results demonstrated that G-DINA and LLM provided the best model fit. While their study focused on test development explicitly grounded in CDM principles, the current research applies a retrofitting approach for diagnostic purposes on pre-existing non-diagnostic tests.

Mohammed et al. (2023) also applied G-DINA to diagnose reading ability using the Preliminary English Test (PET), constructing and validating a Q-matrix reflecting reading comprehension theory and expert consensus. Analysis of 435 test-takers' responses confirmed the utility of G-DINA in generating detailed diagnostic feedback tailored to the PET context. More recently, Boori et al. (2024) compared five constrained CDMs (DINA, DINO, ACDM, LLM, and RRUM) against G-DINA using data from 1,152 examinees on an Iranian high-stakes language proficiency reading test; findings indicated that G-DINA outperformed the other models in terms of model fit and classification accuracy, underscoring its superior ability to capture complex attribute interactions.

2.2. CDMs Applied to This Study

2.2.1. G-DINA Model. In its saturated form, the G-DINA model does not distinguish itself from other general diagnostic models by using different link functions. It accounts for both the main effects and the interaction effects of the attributes. For instance, for an item with two attributes, α_1 and α_2 , the model incorporates one intercept parameter (δ_0 , representing the probability of answering correctly without having mastered any attributes), two main effect parameters ($\delta_1\alpha_1$ and $\delta_2\alpha_2$), and one interaction parameter ($\delta_{12}\alpha_1\alpha_2$). If appropriate restrictions are imposed on these parameters, numerous specific CDMs can be derived from G-DINA (de la Torre, 2011). The probability that test-taker i answers item j correctly is:

$$\rho(x_j=1|\alpha_1\alpha_2) = \delta_{j0} + \delta_{j1}\alpha_1 + \delta_{j2}\alpha_2 + \delta_{j12}\alpha_1\alpha_2$$

2.2.2. DINA Model. The DINA model is a non-compensatory model where an item is answered correctly only if all essential attributes are mastered, allowing for slipping (responding incorrectly despite mastery) and guessing (responding correctly despite lack of mastery). For a two-attribute item, the probability of a correct response is:

$$\rho(x_j=1|\alpha_1\alpha_2) = g_j^{1-\alpha_1\alpha_2} (1 - s_j)^{\alpha_1\alpha_2}$$

Where g_j is the slipping and s_j is the guessing probability. The DINA can be seen as a special case of G-DINA with zero main and interaction effects except for the conjunctive effect (Sijtsma & Junker, 2006).

2.2.3. DINO Model. DINO is a compensatory model in which mastering any one of the necessary attributes is enough to answer the item correctly, while still accounting for slipping and guessing. The probability that test-taker i answers item j correctly can be expressed as:

$$\rho(x_j=1|\alpha_1\alpha_2) = g_j^{(1-\alpha_1)(1-\alpha_2)} (1 - s_j)^{1-(\alpha_1)(\alpha_2)}$$

The probability does not differentiate between mastery of all attributes or any one attribute (Rupp & Templin, 2008; de la Torre, 2011).

2.2.4. A-CDM. The A-CDM is an additive model that originates from the G-DINA by setting all interaction parameters to zero. In this model, each attribute contributes independently to the probability of a correct response, enabling mastery in certain attributes to offset deficiencies in others. It generally employs an identity link and assumes that the overall effect is the sum of each attribute's contribution (de la Torre, 2011). The equation is:

$$\rho(\alpha_{lj}) = \delta_{j0} + \sum_{k=1}^{Kj} \delta_{jk} \alpha_{lk}$$

2.2.5. R-RUM Model. The R-RUM is a non-compensatory CDM that extends the principles of conjunctive attribute mastery. It assumes that the probability of a correct answer is influenced by an item-specific baseline probability and penalty parameters associated with each attribute that has not been mastered. It is regarded as a non-compensatory counterpart of the A-CDM (de la Torre, 2011). The equation is:

$$\rho(\alpha_l) = \delta_{j0} \prod_{k=1}^{Kj} \delta_{jk}^{(1-\alpha_{lk})}$$

To guide this investigation, the following research questions were posed:

- 1) How do the DINO, DINA, A-CDM, LLM, G-DINA, and R-RUM models fit the grammar items of UTEPT at the test level?
- 2) How does the G-DINA model fit the grammar items of UTEPT at the item level?
- 3) What are the weaknesses and strengths of candidates in the grammar section of the UTEPT?

3. Method

3.1. Participants

The participants in the current study (N = 810) were examinees who sat for one of the UTEPT administrations in 2022. This test is a high-stakes standardized assessment administered by the Faculty of Foreign Languages and Literatures at the University of Tehran, primarily required for admission into Ph.D. programs. The sample included 268 males (33.1%) and 542 females (66.9%), all doctoral students in various fields of study, with ages ranging from 24 to 48 years (M = 34.6, SD = 5.8).

3.2. Instruments

The dataset comprised participants' responses to 24 multiple-choice items from the grammar section of the UTEPT, each designed to assess a range of grammatical points. These items were developed based on the item specification table provided by the Language Testing Centre of the Faculty of Foreign Languages and Literatures at the University of Tehran, ensuring alignment with the test's intended construct and overall objectives. The UTEPT consists of several distinct sections: a grammar section includes approximately 40 items, the vocabulary around 30 items, cloze test items near 10, and reading comprehension roughly 20 items, though minor variations exist across administrations.

3.3. Q-matrix Development

In this study, a retrofitting approach was employed to extract diagnostic information from a test that was originally not designed for diagnostic purposes. The initial step in this process was the creation of a Q-matrix, which involves identifying the relevant attributes or sub-skills that underlie test performance. To identify these attributes, researchers can draw from multiple sources, such as theoretical frameworks in the content area, detailed test specifications, think-aloud protocol analyses in which test takers verbalize their cognitive processes, and systematic content analyses of test items (Leighton & Gierl, 2007). Additionally, previous research specific to grammar testing served as an important resource, drawing on studies that investigated grammar sub-skills extensively (Jang, 2008; Langacker, 2009; Lovrić, 2024; Millrood, 2014).

In this study, to identify the attributes of the UTEPT grammar test items, grammar reference books (Azar, 2002; Thomson & Martinet, 2015; Murphy, 2015) were consulted, and prior researches employing CDMs in English grammar testing—where latent attributes and Q-matrices were developed and validated (Liao, 2007; Park & Ho, 2011)—were also reviewed.

A panel of four experienced English language teaching experts analyzed all the test questions. Each grammar item was carefully evaluated to determine the specific grammatical points being assessed. Then, within the framework of defined attributes and based on the opinions of those experts, a list of foundational grammar attributes was prepared based on the research by Park and Cho (2011). Each question was coded according to the attributes required for providing correct answers. The following six latent attributes were defined for the Q-matrix:

Tense Recognition: The ability to identify and understand various verb tenses, which indicate the timing of actions and convey aspects such as continuity or completion (Swan & Walter, 2011).

Voice Awareness: Understanding the distinction between active and passive voice, recognizing when the subject performs the action versus when it receives the action (Azar & Hagen, 2009).

Connectors: The capacity to correctly use conjunctions and transitional phrases that link sentences and ideas, thereby enhancing textual coherence (Linn, 2020).

Clauses: The ability to identify and use different types of clauses—both independent and dependent—that contribute to sentence structure and meaning (Swan & Walter, 2011).

Agreements: Understanding subject-verb and noun-pronoun agreement, including concordance in number and person, and correct pronoun reference (Hewings, 2013).

Lexical knowledge: understanding of words, their meanings, forms, and usages within a language, which is crucial for effective communication and language acquisition (Nation, 2001).

During the Q-matrix development, when experts disagreed, the panel engaged in structured discussions to clarify attribute definitions and resolve item classifications. Ultimately, decisions were made through consensus or majority vote to maintain transparency and ensure the integrity of the Q-matrix. Inter-rater reliability was assessed using Fleiss' Kappa statistic (Fleiss, 1971), which measures the degree of agreement among multiple raters beyond chance. Interpretation of Kappa values follows conventional standards: values less than 0.20 indicate poor agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, 0.61 to 0.80 substantial agreement, and 0.81 to 1.00 almost perfect agreement.

The results demonstrated almost perfect agreement among the raters, particularly for the tense recognition attribute, which exhibited the highest consistency, as shown in Table 1. In the final Q-matrix, as presented in Table 2 (See Appendix A), each row corresponds to a specific item of the UTEPT, while each column represents one of the six latent attributes identified during the study. The values in the matrix are binary, i.e., a value of 1 indicates that the attribute is essential for correctly answering the item, while a value of 0 signifies that the attribute is not required for that particular question.

Table 1
Fleiss' Kappa Coefficient to Assess Interrater Reliability Across the Six Grammar Attributes

Attributes	Kappa Value	Z-Value	P-Value	Agreement Level
Tense Recognition	0.92	9.2	0.0001	Almost perfect agreement
Voice Awareness	0.89	8.9	0.0001	Almost perfect agreement
Connectors	0.87	8.7	0.0001	Almost perfect agreement
Clauses	0.55	5.5	0.0004	Moderate agreement
Lexical knowledge	0.86	8.6	0.0001	Almost perfect agreement

Agreements	0.48	4.8	0.0005	Moderate agreement
------------	------	-----	--------	--------------------

3.4. Procedure

In this study, data analysis was performed using the G-DINA package (version 3.6.3; Ma & de la Torre, 2020) implemented in R software (R Core Team, 2013). This package provides a comprehensive framework for estimating both absolute and relative fit indices, enabling comparison of a model's fit to the data (absolute fit) and comparison among competing models (relative fit). Analyses were conducted at both test and item levels to ensure model validity and diagnostic accuracy. Three relative fit indices guided model evaluation: log-likelihood, which favors higher values when comparing model coefficients; Akaike Information Criterion (AIC; Akaike, 1974); and Bayesian Information Criterion (BIC; Schwarz, 1978). Lower AIC and BIC values indicate better model fit among non-nested models.

Six absolute fit indices were also computed: M^2 , Root Mean Square Error of Approximation (RMSEA²), Standardized Root Mean Square Residual (SRMSR), Proportion Correct (p), Log-Odds Ratio (l), and Transformed Correlation (r). M^2 measures discrepancy between predicted and observed response frequencies, with significant p-values indicating possible item dependency violations (Chen et al., 2025). RMSEA² assesses the discrepancy per degree of freedom between observed and model-implied covariance matrices; values below 0.05 indicate good fit (Maydeu-Olivares & Joe, 2014), while values under 0.06 are acceptable (Hooper et al., 2008). SRMSR reflects the average standardized residuals between observed and expected item correlations, with acceptable ranges of 0 to 0.08 (Hu & Bentler, 1999) and preferable values below 0.05 (Hu et al., 2016).

Residual-based item-level statistics (Xiong et al., 2024) include transformed correlations (r), proportion correct (p), and log-odds ratio (l), measuring discrepancies between observed and model-predicted values. Values near zero, supported by Bonferroni-adjusted p-values greater than 0.05, indicate a good fit. Item-level fit was further examined via (Root Mean Square Error of Approximation) RMSEA, with values below 0.05 indicating excellent fit, 0.05–0.10 moderate fit, and above 0.10 poor fit (Browne & Cudeck, 1992).

Classification accuracy, calculated at both attribute and test levels, assesses agreement between individuals' latent class assignments and their true classes (Cui, Gierl, & Chang, 2012). This study employed G-DINA estimates following Wang et al. (2015). To identify mastery status for each skill, Expected A Posteriori (EAP) estimation was used (Embretson & Reise, 2013). A probability cut-off of 0.5 (Lee & Sawaki, 2009) classified examinees as masters (probability > 0.5) or non-masters (probability ≤ 0.5).

4. Results

4.1. Model Comparison

4.1.1. Model Comparison at the Test Level. The relative and absolute fit indices of the five constrained models compared to the G-DINA model are summarized in Table 3. In terms of the number of parameters, the G-DINA model, with 183 parameters, remains the most complex, while the DINA and DINO models are the most parsimonious, each with 70 parameters. Regarding relative fit indices, the G-DINA model achieved the lowest AIC value (26360.81), indicating the best relative fit when considering model complexity. This was followed by the R-RUM and A-CDM models, with DINA and DINO showing the highest AIC values, reflecting poorer fit. In terms of BIC, which heavily penalizes model complexity, the R-RUM model had the lowest value (26840.87), followed closely by the A-CDM, while the G-DINA ranked higher due to its complexity. As argued by Li et al. (2015), the BIC proposes a large penalty for more highly parameterized models, and this is the main reason why the value of the G-DINA model was high. Absolute fit indices further support the superiority of the G-DINA model. It had the lowest SRMSR (0.0305), well below the acceptable threshold of 0.08, followed by the R-RUM and A-CDM models. The DINA and DINO models exhibited the highest SRMSR values, indicating the worst fit. Similarly, the G-DINA showed the best RMSEA² value (0.0121) with confidence intervals within the good fit criteria, while the R-RUM and A-CDM models also indicated acceptable fit, but DINA and DINO again showed poorer fit. The M^2 statistic indicates adequate fit for the G-DINA model (31.6, $p = 0.235$), whereas all other models indicate misfit. Lastly, the G-DINA

model’s log-likelihood (-13246.41) was the highest, confirming its ability to capture the data complexity better than the constrained models. In general, as the results of the model fit comparison indicate, the G-DINA model demonstrated the best fit among competing models in both relative and absolute terms despite its complexity. The R-RUM and A-CDM offered reasonable alternatives with fewer parameters and acceptable fit, whereas the DINA and DINO models, though more parsimonious, showed the poorest fit across most indices.

Table 3
Relative and Absolute Fit Indices

Models	G-DINA	DINA	DINO	A-CDM	R-RUM
Npars	183	70	70	107	107
AIC	26360.81	26948.30	26958.63	26673.91	26595.96
BIC	26984.95	27106.80	27116.92	26919.06	26840.87
SRMSR	0.0305	0.0549	0.0541	0.0402	0.0402
Log Lik	-13246.41	-13702.10	-13707.16	-13477.85	-13438.83
RMSEA2	0.0121	0.0482	0.0476	0.0369	0.0275
RMSEA2 CI 1	0	0.0437	0.0430	0.0304	0.0224
M2 (p)	31.6 (0.235)	516 (0)	516 (0)	252 (0)	197 (0)

Table 4 presents the absolute item level fit indices for each five models. The results indicate that all models achieved acceptable fit according to the proportion correct values, as the Max[z.stats] for all models were below the critical Z-score of 4.17, and their adjusted p-values equaled 1. This implies that the observed and predicted proportions of correct responses were closely aligned across all models, suggesting no significant misfit at the item level based on this metric. However, the results for the transformed correlations and log-odds ratios reveal notable differences among the models. For these indices, the G-DINA model achieved the most satisfactory results, as its adjusted p-values for both the transformed correlation (0.0678) and log-odds ratio (0.1183) were greater than 0.05, indicating an acceptable fit. In contrast, the DINA, DINO, A-CDM, and R-RUM models showed adjusted p-values below 0.05 for these measures, which is indicative of local item dependencies or model misfit at the pairwise item level.

Table 4
Absolute Item Level Fit Indices

Models		Mean[stats]	Max[stats]	Max[z.stats]	p-value	Adj. p-value
DINA	Proportion correct	0.0011	0.0030	0.2460	0.7762	1
	Transformed correlation	0.0390	0.1465	6.2293	0.0000	0
	Log odds ratio	0.1711	0.7604	5.5322	0.0000	0
DINO	Proportion correct	0.0011	0.0033	0.3092	0.7331	1

	Transformed correlation	0.0400	0.1600	6.8534	0.0000	0
	Log odds ratio	0.1776	0.8502	6.1546	0.0000	0
A-CDM	Proportion correct	0.0017	0.0080	0.7413	0.4600	1
	Transformed correlation	0.0272	0.1126	4.7870	0.0000	0.0021
	Log odds ratio	0.1247	0.5072	4.4865	0.0000	0.0033
R-RUM	Proportion correct	0.0011	0.0050	0.4308	0.6830	1
	Transformed correlation	0.0290	0.1178	4.8163	0.0000	0.0081
	Log odds ratio	0.1253	0.5266	4.5673	0.0000	0.0010
G-DINA	Proportion correct	0.0011	0.0040	0.3653	0.7055	1
	Transformed correlation	0.0247	0.0917	3.7205	0.0000	0.0678
	Log odds ratio	0.1131	0.4415	3.6633	0.0001	0.1183

Note. adj. p-value = adjusted p-value

To evaluate the effectiveness of the different CDMs, classification accuracy was examined at both the attribute level and the test level in accordance with Johnson and Sinharay's (2018) reliability guidelines, which specify thresholds of 0.80 for attribute-level. As Table 5 presents, the attribute-level results reflect excellent reliability for G-DINA and R-RUM across all skills, with occasional weaknesses in DINA and DINO for more complex structures such as Tense Recognition and Voice Awareness.

Table 5
Attribute Level Accuracy

Attributes	G-DINA	A-CDM	R-RUM	DINA	DINO
lexical knowledge	0.922	0.921	0.940	0.916	0.916
Clauses	0.915	0.886	0.925	0.825	0.819
Agreements	0.951	0.891	0.951	0.851	0.848
Connectors	0.948	0.922	0.927	0.867	0.873
Tense Recognition	0.952	0.916	0.948	0.906	0.903
Voice Awareness	0.917	0.903	0.916	0.712	0.870

Table 6 reports the overall test-level accuracy—how well each CDM classified examinees as masters or non-masters of the complete test. Again, applying Johnson and Sinharay's (2018) threshold of 0.70, most models achieved acceptable reliability. The G-DINA model scored the highest, slightly outperforming R-RUM and showing a notable advantage over A-CDM and DINA.

Table 6
Test Level Accuracy

Model	G-DINA	A-CDM	R-RUM	DINA	DINO
Accuracy Index	0.781	0.752	0.780	0.763	0.657

In this section, to evaluate model comparison at the test level, particularly addressing the first question, "How do the DINO, DINA, A-CDM, LLM, G-DINA, and R-RUM models fit the grammar items of UTEPT at the test level?" several key steps were undertaken. First, the Relative Fit Indices were analyzed, revealing that the G-DINA model achieved the lowest AIC (26360.81), indicating the best relative fit among the models. Next, in terms of Absolute Fit Indices, the G-DINA model demonstrated superior performance with the lowest SRMSR (0.0305) and best RMSEA² (0.0121), confirming its overall effectiveness. At the Item-Level, the G-DINA model exhibited the highest adjusted p-values for transformed correlations (0.0678) and log-odds ratios (0.1183), indicating its strong ability to capture item relationships. Additionally, its Attribute Level Accuracy was excellent across various grammar attributes, reinforcing its reliability in assessing attribute mastery. Finally, the G-DINA model achieved the highest test-level accuracy index (0.781), effectively classifying examinees as masters or non-masters. Together, these findings affirm the G-DINA model's superiority in fitting the grammar items of the UTEPT, comprehensively addressing the first Research Question.

4.1.2. Model Comparison at the Item Level

The fit indices in Table 7, serve as a standard for assessing the alignment of the grammar items with the G-DINA model. The RMSEA values for each question indicate how well the items fit the model. Since all RMSEA values for the items in this study are less than 0.05, it can safely be concluded that the grammar items in UTEPT exhibit a good fit with the G-DINA model. Interpreting these results in relation to the second research question, "How does the G-DINA model fit the grammar items of UTEPT at the item level?", the findings confirm that the G-DINA model provides an excellent item-level fit for all grammar items in the test.

Table 7
Item Fit Indices of 24 Items

Item	RMSEA	Item	RMSEA	Item	RMSEA
1	0.003	9	0.001	17	0.004
2	0.004	10	0.005	18	0.002
3	0.005	11	0.002	19	0.003
4	0.002	12	0.004	20	0.004
5	0.004	13	0.003	21	0.002
6	0.007	14	0.003	22	0.005
7	0.003	15	0.004	23	0.001
8	0.006	16	0.002	24	0.003

4.2. Diagnosis of Strengths and Weaknesses of Examinees

To address the third research question, “What are the weaknesses and strengths of candidates in the grammar section of the UTEPT?” the study applied the FAP estimation procedure to determine each examinee’s probability of mastering individual grammar attributes. As presented in Table 8, the frequencies and percentages of examinees in each group were calculated for each skill. The results indicate that 97.0% of examinees mastered the skill of lexical knowledge, making it their strongest area. Conversely, only 18.0% mastered Tense Recognition, which represents the most significant weakness. Other skills, such as voice awareness (34.0%), connectors (38.0%), clauses (44.0%), and agreements (42.0%), also exhibit relatively low mastery levels. These findings suggest that the majority of examinees have not reached mastery in these skills, especially in Tense Recognition and Voice Awareness. The data clearly underscore areas requiring pedagogical attention to strengthen students’ language competencies.

Table 8
The Frequency and Percentage of “Master” and “Non-Master” Persons in Each Attribute

Attributes	Master Frequency	Master Percentage	Non-Master Frequency	Non-Master Percentage
Lexical knowledge	780	97.0%	30	3.0%
Clauses	220	44.0%	590	56.0%
Agreements	210	42.0%	600	58.0%
Connectors	190	38.0%	620	62.0%
Voice Awareness	170	34.0%	640	66.0%
Tense Recognition	90	18.0%	720	82.0%

5. Discussion

This research aimed to develop and validate a Q-matrix and determine the optimal CDM for the grammar section of UTEPT based on the responses of 810 participants. To achieve this, a Q-matrix

was created by identifying the key attributes necessary to answer the grammar questions correctly, inspired by Park and Cho's (2011) classification of grammar attributes, informed by a review of relevant literature, and consultation with four content experts. Six core attributes were specified: Tense Recognition, Voice Awareness, Connectors, Clauses, Agreements, and Lexical Knowledge. The Q-matrix was then constructed and validated with high inter-rater reliability. Subsequently, data were analyzed using five CDMs: DINO, DINA, R-RUM, A-CDM, and G-DINA. Model comparison was conducted at both the test and item levels.

The results address Research Questions 1 and 2 by demonstrating that the G-DINA model provides the best fit for the UTEPT grammar section at both the test and item levels. This aligns with prior studies by Yi (2017), Geramipour et al. (2021), and Shafipour et al. (2021), all of which showed superior performance of the G-DINA model compared to other CDMs in grammar and language proficiency assessments. Additionally, studies by Ravand and Robitzsch (2015), Goodwin et al. (2022), and others consistently find that models capturing complex attribute interactions—such as G-DINA—are optimal for assessing grammar, which involves multifaceted and interdependent skills. The flexibility of the G-DINA as a saturated model is a key strength, as it accommodates both compensatory and non-compensatory relationships among grammar attributes, enabling accurate representation of complex cognitive processes underlying grammar acquisition and use. This flexibility helps provide detailed, accurate feedback that can guide better teaching and learning (Mei & Chen, 2024). In line with the result of Shafipour et al. (2021) research, the findings of this study suggested that the relationships among the attributes of grammar are not ‘either-or’ compensatory or non-compensatory, but a combination of both.

For Research Question 3, examining candidates’ strengths and weaknesses in the grammar section, analysis revealed that Tense Recognition is the most challenging attribute, while lexical knowledge is the easiest. Prior studies align with these findings: Yi (2017) and Shafipour et al. (2021) identified morphological-syntactic knowledge as the most difficult and vocabulary as the simplest. Geramipour et al. (2021) identified verb tense as a weakness and idiomatic expressions as a strength among examinees. This understanding of attribute difficulty offers targeted areas for instructional focus. Moghaddam et al.'s (2016) findings also show that candidates did not achieve mastery in skills related to morphology, conjugation, syntax, and coherence rules, but reached some mastery in vocabulary-related skills.

6. Conclusion

The findings of this study offer theoretical and practical implications for all stakeholders involved in language assessment and teaching. Theoretically, this research contributes to the expanding body of knowledge on retrofitting non-diagnostic tests with CDMs, specifically highlighting the utility of the G-DINA model in providing precise insights into grammar proficiency.

From a pedagogical perspective, the diagnostic feedback from the G-DINA model provides educators with detailed insights into both individual and group mastery of specific grammar attributes. This valuable information enables teachers to tailor instruction, design focused materials, and plan targeted remediation to address the particular weaknesses of learners, thus improving the overall teaching and learning experience (Johnson et al., 2023). The findings from this research indicate that while candidates generally struggle with tense awareness—a challenging grammar skill—they show relatively better mastery in vocabulary-related skills. Because Tense Recognition involves complex grammar rules that are harder to grasp, it is considered a difficult skill. In contrast, vocabulary questions are primarily designed to test sentence structure understanding and are therefore simpler, making vocabulary an easier skill for learners. Consequently, instructional emphasis should be placed more on developing Tense Recognition to better support learner progress.

The study confirms that grammar attributes interact dynamically, reflecting the complex, interdependent cognitive nature of language proficiency. Mastery in some grammar attributes can compensate for weaknesses in others, demonstrating the mental flexibility and interconnectedness of linguistic knowledge. As noted by Shafipour et al. (2021), Geramipour et al. (2021), and Yi (2017), the relationships among grammar attributes are neither purely compensatory nor non-compensatory but combine both patterns. The G-DINA model’s flexible framework accommodates these mixed relationships, aligning well with contemporary cognitive linguistic theories that view grammar as an

integrated, dynamic system rather than isolated skills. This conceptual fit supports the model's ability to capture complex processes underlying grammar acquisition and use, beyond mere statistical accuracy (Croft, 2001; Lakoff, 1986; Langacker, 2008).

Despite the promising results, there are limitations that should be taken into account. The focus on grammar as a single domain may not fully capture the multifaceted nature of language proficiency, which encompasses listening, speaking, and writing skills. Future investigations should consider integrating multiple language domains within diagnostic frameworks to deliver a broader, more integrated assessment of language competence. Exploring the nuances of grammar skills across diverse contexts and combining them with other language skills can lead to more comprehensive and effective diagnostic tools that better inform instructional practices and improve learning outcomes.

Acknowledgments

The authors would like to thank the Faculty of Foreign Languages and Literatures at the University of Tehran and the Language Testing Centre staff for their support and cooperation for providing access to the UTEPT grammar section data used in this study.

Declaration of Conflicting Interests

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

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of AI-Generated Content

AI-assisted tools were not used in this study and the authors remain fully responsible for the content and interpretation of all findings.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Aryadoust, V. (2018). A cognitive diagnostic assessment study of the listening test of the Singapore-Cambridge general certificate of education O-level: Application of DINA, DINO, G-DINA, HO-DINA, and R-RUM. *International Journal of Listening*, 35(1), 29–52. <https://doi.org/10.1080/10904018.2018.1500915>
- Azar, B. (2002). *Understanding and using English grammar*. Pearson Education.
- Azar, B. S., & Hagen, S. A. (2009). *Understanding and using English grammar* (4th ed.). Pearson Longman.
- Barghi, B., Afshinfar, J., Alavi, S. M., Jafarigohar, M., & Soleimani, H. (2025). An exploratory mixed methods study of grammatical range and accuracy in IELTS: A true diagnostic approach to cognitive diagnostic assessment. *International Journal of Language Testing*, 19–42. <https://doi.org/10.22034/ijlt.2024.481434.1374>
- Boori, A. A., Ghazanfari, M., Ghonsooly, B., & Baghaei, P. (2024). Cognitive diagnostic modeling analysis of the reading comprehension section of an Iranian high-stakes language proficiency test. *International Journal of Language Testing*, 14(1), 17–33. <https://doi.org/10.22034/ijlt.2023.399561.1256>
- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2), 230–258. <https://doi.org/10.1177/0049124192021002005>
- 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, Y., Li, X., Liu, J., & Ying, Z. (2025). Item response theory—A statistical framework for educational and psychological measurement. *Statistical Science*, 40(2), 167–194. <https://doi.org/10.1214/23-STS896>

- Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.
- Cui, Y., Gierl, M. J., & Chang, H. H. (2012). Classification accuracy in cognitive diagnostic assessment. *Journal of Educational Measurement*, 49(1), 19–36. <https://doi.org/10.1111/j.1745-3984.2012.00167.x>
- de la Torre, J. (2011). The generalized deterministic inputs, noisy and gate (G-DINA) model. *Psychometrika*, 76(2), 223–251. <https://doi.org/10.1007/s11336-011-9207-7>
- Effatpanah, F., Baghaei, P., & Boori, A. A. (2019). Diagnosing EFL learners' writing ability: A diagnostic classification modeling analysis. *Language Testing in Asia*, 9(1), 12.
- Ellis, R. (2006). Current issues in the teaching of grammar: An SLA perspective. *TESOL Quarterly*, 40(1), 83–107. <https://doi.org/10.2307/40264512>
- Embretson, S. E., & 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>
- Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Psychology Press. <https://doi.org/10.4324/9781410605269>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382. <https://doi.org/10.1037/h0031619>
- 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>
- Goodwin, A. P., Petscher, Y., Tock, J., McFadden, S., Reynolds, D., Lantos, T., & Jones, S. (2022). Monster, PI: Validation evidence for an assessment of adolescent language that assesses vocabulary knowledge, morphological knowledge, and syntactical awareness. *Assessment for Effective Intervention*, 47(2), 89–100. <https://doi.org/10.1177/1534508420966383>
- Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, 32(3), 317–336. <https://doi.org/10.1177/0265532214564505>
- Hemmat, S. J., Baghaei, P., & Bemani, M. (2016). Cognitive diagnostic modeling of L2 reading comprehension ability: Providing feedback on the reading performance of Iranian candidates for the university entrance examination. *International Journal of Language Testing*, 6(2), 92–100.
- Hewings, M. (2013). *Advanced grammar in use with answers: A self-study reference and practice book for advanced learners of English*. Cambridge University Press.
- Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, 6(1), 53–60.
- Hu, J., Miller, M. D., Huggins-Manley, A. C., & Chen, Y. H. (2016). Evaluation of model fit in cognitive diagnosis models. *International Journal of Testing*, 16(2), 119–141. <https://doi.org/10.1080/15305058.2015.1133627>
- Hu, L., & Bentler, P. M. (1999). Cut off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 117–131).
- Jang, E. E. (2009). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly*, 6(3), 210–238. <https://doi.org/10.1080/15434300903071817>
- Javidanmehr, Z., & Anani Sarab, M. R. (2019). Retrofitting non-diagnostic reading comprehension assessment: Application of the G-DINA model to a high-stakes reading comprehension test. *Language Assessment Quarterly*, 16(3), 294–311. <https://doi.org/10.1080/15434303.2019.1654479>
- Jiao, H. (2009). Diagnostic classification models: Which one should I use. *Measurement: Interdisciplinary Research & Perspective*, 7(1), 65–67. <https://doi.org/10.1080/15366360902799869>

- Johnson, C. C., Walton, J. B., Strickler, L., & Elliott, J. B. (2023). Online teaching in K–12 education in the United States: A systematic review. *Review of Educational Research*, 93(3), 353–411. <https://doi.org/10.3102/00346543221105550>
- Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement*, 55(4), 635–664. <https://doi.org/10.1111/jedm.12196>
- Kim, Y. H. (2011). Diagnosing EAP writing ability using the reduced parameterized unified model. *Language Testing*, 28, 509–541. <https://doi.org/10.1177/0265532211400860>
- Lakoff, G. (1986). Frame semantic control of the coordinate structure constraint. In A. M. Farley, P. T. Farley, & K.-E. McCullough (Eds.), *Papers from the parasession on pragmatics and grammatical theory* (pp. 152–167). Chicago Linguistic Society.
- Langacker, R. (2008). *Cognitive grammar: A basic introduction*. Oxford University Press.
- Langacker, R. W. (2009). Cognitive grammar. In *Cognition and pragmatics* (pp. 78–85). John Benjamins Publishing Company. <https://doi.org/10.1017/S0022226709005799>
- Lee, Y.-W., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly*, 6, 172–189. <https://doi.org/10.1080/15434300902985108>
- Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.
- Li, H., Hunter, C. V., & Lei, P. W. (2015). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, 33(3), 391–409. <https://doi.org/10.1177/0265532215590848>
- Liao, Y. F. (2007). Investigating the construct validity of the grammar and vocabulary section and the listening section of the ECCE: Lexico-grammatical ability as a predictor of L2 listening ability. *Spain Fellow Working Papers in Second or Foreign Language Assessment*, 5, 37–60.
- Linn, A. R. (2020). English grammar writing. In *The handbook of English linguistics* (pp. 63–80). <https://doi.org/10.1002/9781119540618.ch5>
- Lovrić, D. (2024). *Grammar assessment in EFL classrooms* [Doctoral dissertation, University of Zadar, Department of English].
- Ma, W., & de la Torre, J. (2020). G-DINA: An R package for cognitive diagnostic modeling. *Journal of Statistical Software*, 93(14), 1–26. <https://doi.org/10.1007/s11336-011-9207-7>
- Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, 49(4), 305–328. <https://doi.org/10.1080/00273171.2014.911075>
- Mei, H., & Chen, H. (2024). Cognitive diagnosis in language assessment: A thematic review. *RELC Journal*, 55(2), 529–537. <https://doi.org/10.1177/00336882221122357>
- Millrood, R. (2014). Cognitive models of grammatical competence of students. *Procedia-Social and Behavioral Sciences*, 154, 259–262. <https://doi.org/10.1016/j.sbspro.2014.10.147>
- Mirzaei, A., Vinchek, 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, 100817. <https://doi.org/10.1016/j.stueduc.2019.100817>
- Moghaddam, A., Falsafinejhad, M., Farokhi, N., & Estaji, M. (2016). Diagnostic analysis of general English reading comprehension's items of PhD entrance exam using non-compensatory fusion model. *Quarterly of Educational Measurement*, 22, 41–68. [in Persian]
- Mohammed, A., Dawood, A. K. S., Alghazali, T., Kadhim, Q. K., 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), 1–20.
- Murphy, R. (2015). *Essential grammar in use: A self-study reference and practice book for elementary students of English; with answers*. Cambridge University Press.
- Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press. <https://doi.org/10.1017/9781009093873>
- Park, C., & Cho, S. (2011). Cognitive diagnostic assessment of English grammar for Korean EFL learners. *English Teaching*, 66(4), 101–130. <https://doi.org/10.15858/engtea.66.4.201112.101>

- R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <http://www.r-project.org/>
- Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research, and Evaluation*, 20(1). <https://doi.org/10.7275/5g6f-ak15>
- 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>
- Ravand, H., Effatpanah, F., Ma, W., de la Torre, J., Baghaei, P., & Kunina-Habenicht, O. (2024). Exploring interrelationships among L2 writing subskills: Insights from cognitive diagnostic models. *Applied Measurement in Education*, 37(4), 329–355. <https://doi.org/10.1080/08957347.2024.2424550>
- Rupp, A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, 6, 219–262. <https://doi.org/10.1080/15366360802490866>
- Schwarz G (1978) Estimating the dimension of a model. *Annals Stat* 6(2). <https://doi.org/10.1214/aos/1176344136>
- Shafipour, M., Ravand, H., & Maftoon, P. (2021). Test-level and item-level model fit comparison of general vs. specific diagnostic classification models: A case of true DCM. *Language Testing in Asia*, 11, 1–20. <https://doi.org/10.1186/s40468-021-00148-z>
- Sijtsma, K., & Junker, B. W. (2006). Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika*, 33(1), 75–102.
- Swan, M., & Walter, C. (2011). *Oxford English grammar course: Advanced*. Oxford University Press. <https://doi.org/10.1093/elt/ccs003>
- Thomson, A. J., & Martinet, A. V. (2015). *A practical English grammar*. Oxford University Press.
- VanPatten, B. (2017). Grammar learning and processing: Building on the processability model. *Language Learning*, 67(S1), 161–182.
- Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, 52(4), 457–476. <https://doi.org/10.1111/jedm.12096>
- Xie, Q., & Lei, Y. (2021). Diagnostic assessment of L2 academic writing product, process and self-regulatory strategy use with a comparative dimension. *Language Assessment Quarterly*, 19(3), 231–263. <https://doi.org/10.1080/15434303.2021.1903470>
- Xiong, J., Luo, Z., Luo, G., Yu, X., & Li, Y. (2024). An exploratory Q-matrix estimation method based on sparse non-negative matrix factorization. *Behavior Research Methods*, 56(7), 7647–7673.
- 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. <https://doi.org/10.1080/08957347.2017.1283314>

Appendix A

The finalized Q-matrix

The finalized Q-matrix in Appendix A shows the relationships between each test item and six grammar attributes, indicating which specific skills are required to answer each item correctly.

Table 2

Final Q-matrix for the Grammar Test

Item	A1	A2	A3	A4	A5	A6	Item	A1	A2	A3	A4	A5	A6
1	1	1	0	0	1	0	13	0	0	0	1	0	0
2	0	0	1	0	0	1	14	0	0	0	0	1	0
3	0	0	0	0	0	1	15	0	0	1	0	0	0
4	0	0	0	1	0	1	16	0	0	0	1	0	1
5	0	0	1	0	0	0	17	1	0	0	0	0	1
6	0	1	0	0	0	1	18	0	0	1	0	0	1
7	0	0	0	0	0	1	19	1	0	1	0	0	1
8	0	1	0	0	0	0	20	1	0	0	0	0	1
9	0	0	0	0	1	0	21	0	0	0	0	0	1
10	0	0	0	0	0	1	22	1	0	0	0	0	1
11	0	0	0	0	0	1	23	0	0	1	0	0	0
12	1	0	0	0	0	0	24	0	0	0	0	0	1

Note. A1 = Tense Recognition, A2 = Voice Awareness, A3 = Connectors, A4 = Clauses, A5 = Agreements, A6 = lexical knowledge. 1 = Attribute required; 0 = Attribute not required.