Tabaran Institute of Higher Education
*International Journal of Language Testing*
Vol. 15, No. 1; 2025, March

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

# Constructing and Validating a Q-matrix for Cognitive Diagnostic Analysis of the Listening Comprehension Section of the IELTS

Seyedeh Azadeh Ghiasian[1], Fatemeh Hemmati[2]*, Seyyed Mohammad Alavi[3], Afsar Rouhi[4]

**ARTICLE INFO**

**Article History:**
Received: May 2024
Accepted: June 2024

**KEYWORDS**
Attributes
CDMs
IELTS
Listening Comprehension
Q-matrix

**ABSTRACT**

A critical component of cognitive diagnostic models (CDMs) is a Q-matrix that stipulates associations between items of a test and their required attributes. The present study aims to develop and empirically validate a Q-matrix for the listening comprehension section of the International English Language Testing System (IELTS). To this end, a listening comprehension test of the IELTS was administered to 820 Iranian test takers. According to theories, taxonomies, and models of second/foreign language (L2) listening comprehension, previous studies on the utility of CDMs to L2 listening comprehension, detailed content analysis of the test items, and consultation with several content experts, an initial Q-matrix was first developed. Through the technique suggested by de la Torre and Chiu (2016), along with checking heatmap plots and mesa plots using the GDINA package in R, the Q-matrix was then empirically validated. Generally, six attributes were extracted for the listening section, namely, (1) Linguistic knowledge (LKA), (2) understanding prosodic patterns (UPP), (3) ability to understand and make paraphrases (PAR), (4) ability to understand specific factual information such as names, numbers, and so forth (UFI), (5) ability to understand explicit information (UEI), and (6) ability to make inference (INF). Finally, the results of the fit of the GDINA model to the data, at both item and test levels, indicated the adequate model-data fit and the plausibility of the Q-matrix. The implications of the study were also discussed.

## 1. Introduction

In contemporary language assessment, the International English Language Testing System (IELTS) serves as a pivotal instrument for evaluating proficiency in English, often for purposes of immigration or academic pursuits abroad (Nakatsuhara et al., 2017; Phakiti, 2016). Among its various components, the listening comprehension section of the IELTS stands out as a critical evaluation tool. Listening comprehension, a multifaceted process crucial in language acquisition, involves intricate cognitive mechanisms and linguistic knowledge (Snowling & Hulme, 2005).

Despite the widespread adoption of the IELTS, scholarly attention directed specifically towards its listening comprehension section has been limited (e.g., Aryadoust, 2011, 2012, 2013; Badger & Yan, 2009; Field, 2012; Harding et al., 2015; Phakiti, 2016; Winke & Lim, 2014). Consequently, there exists a notable research gap in understanding the underlying processes and attributes essential for the successful completion of this section. Moreover, existing educational programs aimed at preparing test

[1] Department of English Language Teaching, Payame Noor University, Tehran, Iran, Email: ghiasian63@student.pnu.ac.ir
[2] Department of English Language Teaching, Payame Noor University, Tehran, Iran, Email: hemmati@pnu.ac.ir
[3] University of Tehran, Email: smalavi@ut.ac.ir
[4] University of Mohaghegh Ardabili, Ardabil, Email: afsarrouhi@uma.ac.ir

Cite this paper as: Ghiasian, A., Hemmati, F., Alavi, M., & Rouhi, A. (2025). Constructing and validating a Q-matrix for cognitive diagnostic analysis of the listening comprehension section of the IELTS. *International Journal of Language Testing, 15*(1), 54–76. https://doi.org/10.22034/IJLT.2024.456609.1344

takers often lack diagnostic precision, offering insufficient feedback to identify individual weaknesses and strengths in cognitive domains pertinent to listening comprehension (Aryadoust, 2012).

Following this logic in educational contexts, cognitive diagnostic assessment (CDA) serves as a valuable approach to precisely evaluate learners' proficiency in particular skills or attributes (Chen et al., 2013). While traditional psychometric frameworks offer valuable insights, cognitive diagnostic models (CDMs) provide a nuanced understanding by dissecting test items into specific attributes or skills required for proficiency (Embretson, 1983). Amid the evolving landscape of language testing and assessment, the significance of cognitive diagnosis, particularly in standardized high-stakes testing such as the IELTS, becomes increasingly apparent (de la Torre, 2009). By employing CDA within the framework of the IELTS listening comprehension section, educators can gain insights into individual test takers' strengths and weaknesses, thereby facilitating targeted instructional design and test preparation strategies (Min & He, 2021).

Utilizing CDMs as valuable means to delve into the cognitive processes and fundamental attributes associated with various skills is facilitated notably by the incorporation of a Q-matrix (Andringa et al., 2012). The Q-matrix, a foundational element within CDMs, elucidates the relationships between test items and the specific attributes they evaluate, using binary indicators (1 and 0) to denote the presence or absence of each attribute required for correct item response. Numerous methodologies have been proposed by researchers to delineate the attributes pertinent to constructing a Q-matrix for a given test. These methods encompass diverse approaches such as content analysis of test items, examination of established test specifications, application of content domain theories, comprehensive literature review, consultation with expert panels, dimensionality analysis, eye-tracking research, and think-aloud protocols (Gao & Rogers, 2007; Jang, 2009; Sawaki et al., 2009). Scholarly evidence indicates that precise specification of underlying attributes associated with a set of test items or tasks, along with their theoretically-grounded relationships with items, enhances the efficacy of CDA (Lee & Sawaki, 2009a).

However, despite the array of strategies available for attribute determination, the process of Q-matrix construction often entails a degree of subjective judgment. This subjectivity inherent in Q-matrix development may predispose to misspecifications, which can impact model parameters, classification accuracy, and ultimately compromise the validity of inferences drawn (Chiu, 2013; de la Torre & Chiu, 2016; Madison & Bradshaw, 2015). Consequently, various methods for Q-matrix validation have been devised to detect and rectify such misspecifications (e.g., de la Torre et al., 2022; Kang et al., 2019; Li et al., 2021; Ma & de la Torre, 2020; Nájera et al., 2020; Wang et al., 2018). Among these, the approach proposed by de la Torre and Chiu (2016) emerges as particularly prominent. This method involves initially employing the Generalized-Deterministic Input, Noisy and Gate (G-DINA) model as a foundational framework for the dataset. The G-DINA model has gained attention for its potential in assessing language proficiency and identifying specific areas of strength and weakness in test takers (Harding et al., 2015; Min & He, 2021).

Considering the high-stakes nature of the IELTS exam and its substantial implications for test takers, the integration of cognitive diagnostic modeling holds significant promise for enhancing the diagnostic accuracy of the test. By identifying mastery and non-mastery of specific subskills, educators and test developers can tailor interventions and instructional strategies to address individual learning needs more effectively. Given this context, this study aims to develop and validate a Q-matrix for the listening comprehension section of the IELTS, considering its status as a high-stakes examination. The study posed the following research questions:

**RQ1.** What are the primary underlying L2 listening processes or attributes essential for effectively completing the listening comprehension segment of the IELTS?

**RQ2.** Does the G-DINA model have a sufficient fit to the listening section of the IELTS using the final validated Q-matrix?

This research builds on previous studies applying CDMs to listening comprehension by utilizing a larger and more representative sample, ensuring the production of robust and generalizable results. The study advances the ongoing development and refinement of Q-matrices for standardized language proficiency tests through several methodological innovations. These include the use of mesa plots to verify the plausibility of *q*-vectors, a technique rarely used in earlier research, and the employment of heatmap plots to assess item pair dependencies, an innovative approach largely

overlooked in prior CDM applications. The study contributes to the enhancement of diagnostic precision in language assessment, thereby facilitating targeted instructional interventions for trainers and also language learners preparing for the IELTS examination.

## 2. Review of Literature

### *2.1. Listening Comprehension*

Listening comprehension in second language learning is a multifaceted process that entails intricate cognitive mechanisms and diverse theoretical perspectives. The literature on this topic delineates two primary groups of listening models, each offering unique insights into the nature of listening and comprehension (Rost, 2016; Vandergrift & Goh, 2012). The initial category, as portrayed by academics like Vandergrift and Goh (2012) and Rost (2016), consists of broad frameworks concentrating on listening in settings without assessment. Conversely, the subsequent category includes models specifically designed for assessment, such as the listening-response model introduced by Bejar et al. (2000). These models delineate listening as a dual-phase process, encompassing comprehension during the listening phase and verbal or nonverbal reactions during the response phase.

Within the assessment-specific models, the default listening construct emphasizes the significance of authentic stimuli and encompasses bottom-up and top-down comprehension processes (Buck, 2001). Additionally, research-based frameworks, validated through quantitative analysis, contribute to a deeper understanding of listening proficiency (Buck & Tatsuoka, 1998; Liao, 2007; Sawaki et al., 2009). On the other hand, non-assessment listening models regard listening as a multifaceted cognitive process involving both pre-comprehension and comprehension stages (Dunkel et al., 1993). The pre-comprehension phase involves perception and recognition mechanisms, including lexical segmentation and access, while comprehension encompasses selection or construction processes guided by syntactic knowledge (Kintsch, 1998).

Assessment-specific models identify specific listening subskills, including understanding details, key vocabulary, and speaker attitudes (Sawaki et al., 2009; Vandergrift, 2007; Wolfgramm et al., 2016). Furthermore, these models acknowledge that factors related to testing, such as content, context, and the inclusion of visual cues, can impact listening ability (Bejar et al., 2000). Facets such as eliminating distractors and understanding main ideas can significantly impact test takers' performance (Ackerman et al., 2005). Moreover, the concept of "extra-listening" encompasses non-listening construct-irrelevant processes, such as concurrent reading, which may affect test validity (Ackerman et al., 2005). Drawing from a multitude of theories, the body of literature concerning listening comprehension in second language acquisition presents a wide array of theoretical frameworks and empirical observations. These contributions illuminate the intricate cognitive processes and assessment considerations associated with this fundamental aspect of language acquisition.

### *2.2. Cognitive Diagnostic Models*

Recently, there has been a notable increase in interest regarding CDMs due to their ability to provide intricate insights into students' learning status, thus facilitating the implementation of targeted instructional strategies (Rupp et al., 2010). Unlike traditional psychometric frameworks like classical test theory (CTT) and item response theory (IRT), which typically concentrate on a single proficiency continuum epitomized by a true score or latent trait, CDMs present a more nuanced approach by offering comprehensive diagnostic feedback regarding a student's cognitive abilities' strengths and weaknesses (Lee et al., 2012). These models operate under the assumption of the involvement of multiple cognitive skills, strategies, and knowledge domains essential for accurate responses to test items or tasks (Birenbaum et al., 1993). This inherent characteristic of CDMs enables them to yield multidimensional diagnostic profiles based on statistically-driven multivariate classifications (Rupp & Templin, 2008) that outline test takers' mastery levels across various traits. The insights derived from these profiles can then be utilized to tailor remedial interventions and enhance instructional approaches. According to Rupp and Templin (2008), CDMs are:

> probabilistic, confirmatory multidimensional latent-variable models with a simple or complex loading structure. They are suitable for modeling observable categorical response variables and contain unobservable (i.e., latent) categorical

> predictor variables. The predictor variables are combined in compensatory and non- compensatory ways to generate latent classes. (p. 226)

Numerous CDMs have been created and developed, including the Deterministic Inputs, Noisy "or" Gate (DINO; Templin & Henson, 2006), the Deterministic Inputs, Noisy "and" Gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001), the Linear Logistic Model (LLM; Maris, 1999), the Additive CDM (*A*-CDM; de la Torre, 2011), the Reduced Reparameterized Unified Model (RRUM or Fusion Model; Hartz, 2002), the Log-linear Cognitive Diagnosis Model (LCDM' Henson et al., 2008), the General Diagnostic Model (GDM; von Davier, 2008), and the Generalized Deterministic, Inputs, Noisy "and" Gate (G-DINA' de la Torre, 2011).

One of the important general models is the G-DINA model. Similar to any other general CDM, all potential main and interaction effects are taken into account, and various interactions across attributes (such as compensatory and non-compensatory interactions) are allowed (de la Torre, 2011). One of the distinctive features of the model is its departure from the conjunctive assumption of the DINA model, which categorizes examinees into 2 groups for each item. The G-DINA model categorizes test-takers into $2^{k_j^*}$ latent groups, where $k_j^*$ is the total number of attributes needed to complete item *j*. For a test taker with an attribute pattern $\alpha_{lj}^*$, the likelihood of providing an accurate response depends on the primary effects and all potential interaction effects among the $k_j^*$ necessary abilities for item *j*:

$$(1) \quad P(\alpha_{lj}^*) = \delta_{j0} + \sum_{k=1}^{k_j^*} \delta_{jk}\, \alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{k_j^*-1} \delta_{jkk'} \alpha_{lk} a_{lk'} \ldots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}$$

where $\delta_{j0}$ represents the intercept for item *j* (e.g., the probability of a correct response when none of the required skills is present); $\delta_{jk}$ denotes the main effect due to a single attribute $\alpha_k$, indicating the change in the probability of success upon mastering a single attribute (i.e., $a_k$); $\delta_{jkk'}$ signifies the (first-order) interaction effect between $\alpha_k$ and $\alpha_{k'}$ reflecting the change in the probability of success resulting from the mastery of both $\alpha_k$ and $\alpha_{k'}$; $\delta_{j12\ldots k_j^*}$ represents the highest-order interaction effect attributed to $\alpha_{1,\ \ldots,}\alpha_{k_j^*}$ indicating the probability of a correct response due to the mastery of all the required skills, which exceeds the additive impact of all the main lower-order interaction effects (de la Torre, 2011).

### *2.3. Previous Applications of CDMs to Listening Comprehension*

Although researchers have increasingly employed CDMs in various language proficiency tests, previous studies predominantly focused on reading (e.g., Boori et al., 2023; Boori et al. 2024; Chen et al., 2022; Chen & Chen, 2016; Ranjbaran & Alavi, 2017; Ravand, 2016) and writing (Effatpanah et al., 2019; Kim, 2014; Xie, 2016) skills. In the area of listening comprehension, Buck and Tatsuoka (1998) conducted a seminal study, utilizing the Rule Space Methodology to delineate the linguistic and cognitive attributes underpinning a free-response listening test. Attributes were derived through a detailed analysis of test items and a literature search, resulting in a Q-matrix that identified 15 primary attributes and 14 interaction attributes. They concluded that the rule space methodology can be used to accurately classify test takers into different latent knowledge states. However, the complexity and extensive data requirements of the Rule Space Model may limit its practical application in broader testing contexts.

Lee and Sawaki (2009a) conducted a comprehensive multi-CDM research on the listening and reading sections of the IBT TOEFL to explore the cognitive attributes underlying test performance, comparing the performance of three cognitive diagnostic models (i.e., Fusion model, GDM, and latent class analysis model). Attributes were derived through a content analysis of individual test items conducted by content experts for the development of the Q-matrices. The results demonstrated that all three cognitive diagnosis models were effective in identifying the underlying cognitive attributes of ESL reading and listening assessments. Their analysis demonstrated comparable performance across these models, concerning probabilities of skill mastery, classification of test-takers' skills, and the reliability of classification. The study also found variability in the precision and consistency of the attribute classifications across the different models.

Aryadoust (2018) investigated five CDMs (including DINA, DINO, G-DINA, HO-DINA, and RRUM) to unveil the fundamental framework of the listening test in the Singapore-Cambridge General Certificate of Education (GCE) exam. In this study, the creation of the Q-matrix was influenced by three primary sources of information: the theoretical framework guiding the research, think-aloud data obtained from interviews with 20 participants, and insights gathered from eye-tracking technology. His analysis uncovered that the nine subskill RRUM demonstrated superior fitting in comparison to other models, indicating the challenging nature of certain listening attributes for test-takers to master.

In a study, Alavi et al (2018) investigated the construct validity of the IELTS listening comprehension test with the use of structural equation modelling (SEM) and assessed differential item functioning (DIF) through CDMs and Mantel Haenszel (MH). The results indicated that the listening module of IELTS possesses substantial validity, supported by consistent evidence across various sources. They also found that the four constructs of diagram labeling, gap filling, short answer, and multiple-choice within the IELTS listening comprehension test framework significantly contributed to its validity. Furthermore, their study reported that MH identified 15 DIF items, while CDM identified between 6 and 12 DIF items.

Effatpanah (2019) compared the performance of five cognitive diagnostic models (including DINA, NC-RUM, ACDM, DINO, and C-RUM) against the G-DINA model to delineate the inherent interplay among listening attributes in the listening section of the IELTS exam and diagnose listening ability of Iranian examinees. Attributes were derived through a detailed process involving expert judgment, literature review, and content analysis of the test items. The findings indicated that among the competing models, the G-DINA model demonstrated superior fit across all indices. The C-RUM model, having the closest performance to the G-DINA model, emerged as the best specific CDM among competing models. It was revealed that inference-making, vocabulary, and syntax were identified as particularly challenging for Iranian candidates.

Dong et al. (2021) explored CDM selection for an L2 listening comprehension test (L2LDA) within the PELDiaG system, employing statistical and content analyses to confirm optimal model choices and attribute relationships. The researchers evaluated several CDMs (including the DINA, DINO, R-RUM, A-CDM, and LLM and G-DINA models) to determine which provided the best fit for the listening comprehension test data. Attributes were derived through the study of Meng (2013). At the test level, the A-CDM, LLM, and R-RUM demonstrated acceptable and comparable model fit, indicating mixed inter-attribute relationships among L2 listening subskills. At the item level, Mixed-CDMs were preferred and validated, confirming the presence of these mixed relationships. Mixed-CDMs outperformed the G-DINA model in terms of model and person fit. Alongside statistical methods, content analysis offered theoretical support to validate and refine the item-level CDMs. It was also noted that sample size and the number of multi-attribute items are critical considerations in L2 listening cognitive diagnostic modeling studies.

These studies collectively highlight the potential of CDMs in identifying strengths and weaknesses of examinees, offering valuable tools for educational assessment and instructional improvement. However, they have several limitations regarding Q-matrix validation. For instance, Aryadoust (2018) did not conduct any empirical Q-matrix validation. Except for Effatpanah (2019), the majority of other studies relied on non-compensatory models, such as the Fusion Model, for Q-matrix development. This procedure might influence the results (Lee & Sawaki, 2009b). If such a Q-matrix, refined using a non-compensatory model, is used for model comparison at test-level, the non-compensatory models might have been given an advantage over the compensatory models (Ravand & Robitzsch, 2018). Additionally, some studies, such as Lee and Sawaki (2009b), utilized three different software programs to estimate the models.

Furthermore, too little attention has been devoted to the application of CDMs to listening comprehension section of large-scale language proficiency assessments, especially IELTS. Effatpanah's (2019) study, the only attempt to apply CDMs to listening comprehension, faced a notable limitation: a relatively small sample size. This constraint is critically important, as small sample sizes can potentially undermine the reliability of CDM outcomes, such as classification accuracy and mastery profiles (Tatsuoka, 2009). Additionally, CDMs are sensitive to the structure of Q-matrices. Different Q-matrices can yield different interpretations and conclusions, underscoring the imperative for rigorous

validation procedures (de la Torre & Chiu, 2016). Thus, further research efforts are warranted to refine and develop Q-matrices tailored to standardized language proficiency tests.

## 3. Method

### *3.1. Participants and Setting*

The participants consist of 820 EFL (English as a Foreign Language) learners from various universities and IELTS candidates from language institutes in Iran. Their scores ranged from 11 to 40. Out of the entire sample, 532 individuals (64.8%) were female, while 288 individuals (35.1%) were male. The ages of the participants ranged from 18 to 36 years, with an average age of 21.4 years (SD= 3.9).

Furthermore, the development stage of the Q-matrix involved the collaboration of the researchers and four experienced IELTS instructors as an expert panel. They were all non-native English speakers whose primary language was Persian, with English being their second language. Among them were three IELTS instructors boasting over a decade of experience in teaching both general English and IELTS, alongside an educational supervisor with approximately 20 years of expertise in English instruction and global high-stakes examinations, and two university professors with nearly 30 years of teaching English as a foreign language. Each instructor held either a Ph.D. or Master's degree in TEFL (Teaching English as a Foreign Language). Their ages spanned from 36 to 65 years old.

### *3.2. Instrumentation*

The investigation employed the listening section of an IELTS examination, comprising four sections, each featuring 10 inquiries. Participants were allocated time to review instructions and questions, and also to review their work after each section. All recordings were played once. The initial task required participants to listen to a conversation where a man contacts a catering company to arrange a party. Participants were tasked with carefully listening to the dialogue, completing a customer booking form based on the provided information, and identifying final decisions. This section consisted of two multiple-choice and eight fill-in-the-gap items.

In the subsequent task, participants were presented with a guide's explanations to a group of tourists regarding Buckingham Palace and its historical changes. Seven questions inquired about events related to the Palace, and participants were tasked with determining the timing of these events by selecting from three date period options provided at the beginning. The remaining three questions required participants to fill in gaps with no more than two words and/or a number, as per the instructions.

In the third section, examinees listened to a conversation between two university students discussing a social science lecture they attended. Four questions required responses of no more than three words, as specified, while the other six questions were matching items concerning the lecture's content. Finally, in section four, participants listened to a lecture by a professor on a specific period in American history. This part included three fill-in-the-gap and seven multiple-choice items. Following the test, participants were granted ten minutes to transfer their answers to answer sheets**.** The reliability coefficient of the listening test was estimated using Cronbach's alpha, and a value of 0.87 was obtained, which is widely acknowledged.

### *3.3. Procedures*

The duration of collecting data for the study spanned 12 weeks, during which ethical standards, including informed consent, respect for anonymity, and confidentiality, were strictly adhered to. To safeguard participant anonymity, all student names were omitted from the collected data. After preparing the data, various techniques were employed to determine the attributes or cognitive processes necessary for test takers to successfully answer test items. It has been demonstrated that developing a Q-matrix from multiple sources maximizes the Q-matrix's consistency (Li & Suen, 2013).

The methodology employed in this study involved several detailed steps. Initially, the researchers conducted a comprehensive review of relevant literature and theories to compile a list of attributes (e.g., Buck, 2001; Buck & Tatsuoka, 1998; Field, 2009; Flowerdew & Miller, 2005; Goh, 2000; Rost, 2016; Yi, 2017). For instance, Richards's (1983) lists of micro-skills include 33 micro-skills for "conversational listening" and 18 micro-skills for "academic listening". Expanding upon Richards's (1983) categorization of aural skills, Brown (2007) presents a streamlined compilation of

micro-skills and macro-skills, particularly focusing on conversational listening. He distinguished 10 listening comprehension micro skills and 7 macro skills for conversational discourse.

For the second source, the researchers utilized previous research conducted on L2 listening comprehension within the context of cognitive diagnostic modeling. Table 1 summarizes a compilation of listening attributes derived from some prior studies. While certain attributes are unlikely to directly relate to the test utilized in the present study, they still aid in the selection of pertinent and applicable attributes.

**Table 1**

*Summary of Listening Attributes Recognized in Some of the Prior Studies on L2 Listening Comprehension*

| **Studies** | **Extracted Attributes** |
|---|---|
| Lee and Sawaki (2009) (the listening section of the TOEFL IBT) | "- Understanding General Information,<br>- Understanding Specific Information,<br>- Understanding Text Structure and Speaker Intention,<br>- Connecting Ideas" (p. 246) |
| Aryadoust (2012) (through content analysis of the listening section of IELTS) | "- Linguistic repertoire<br>- World knowledge sources (schema)<br>- Ability to make paraphrases<br>- Ability to understand specific factual information such as names, numbers, and so forth<br>- Integrate listening ability and visual skills<br>- Integrate listening, reading, short-term memory span, and/or writing abilities" (p. 60) |
| Effatpanah (2019) (the listening section of IELTS) | "- Making inferences (Tsui & Fullilove, 1998);<br>- Understanding paraphrases (PAR) (Wagner, 2004);<br>- Understanding detailed information (DET) (Sawaki et al., 2009);<br>- Understanding explicitly stated general and literal information (LIT) (Field, 2009);<br>- Comprehending vocabulary and syntax (VOG) (Aitken, 1978; Shin, 2008; Wolfgram et al., 2016);<br>- Keeping up with the pace of speakers (PAC) (Richards, 1983);<br>- Identifying prosodic patterns and speakers' attitudes and intentions (PPS) (Aitken, 1978; Vandergrift, 2007)" (p. 10) |
| Dong et al (2021) (L2LDA as part of the English as Foreign Language Listening Diagnostic Test in the PELDiaG system [Personalized English Learning Diagnosis and Guidance system]) | "- Sound Discrimination: Recognizing special phonological and prosodic information, such as liaison and assimilation, stress and weak forms, intonation;<br>- Less Frequent Vocabulary and Expressions: Understanding less frequent words, oral expressions, and slangs;<br>- Difficult Structures: Difficult sentence structure and grammatical functions such as subjunctive mood, inversion, and negation;<br>- Facts and Details: Understanding detailed expressions of time, places, and relationships;<br>- Main Idea: Recognizing and summarizing main ideas and major points;<br>- Situational Context and Cultural Background Inferences: Obtaining motivations, purposes, reasons, and interactive functions by inferring from the context, implied expressions, and cultural background" (p.4) |

Aryadoust (2018, p. 4) also outlined the enumeration of primary sub-skills that have been identified and validated through various listening assessments:

> (1) Understanding details (Sawaki et al., 2009); (2) Understanding key vocabulary (Wolfgramm et al., 2016); (3) Making paraphrases to connect the listening text to the test items (Wagner, 2004); (4) Obtaining inferences, e.g., propositional (logical conclusions based on the facts in the listening text), enabling (linking the listening input to one's own world knowledge) (Wagner, 2004; drawn from Hildyard & Olson, 1978); (5) Differentiating main ideas from details (Yeldham, 2016); (6) Understanding speakers' attitudes and intentions through prosody and vocabulary (Vandergrift, 2007); (7) Drawing conclusions (Liao, 2007).

After examining the relevant literature and previous CDM studies on L2 listening, a list of attributes was compiled. This foundational list was then refined through collaborative brainstorming sessions between the researchers and four additional content experts. These sessions aimed to delineate the correlation between each test item and the attributes it necessitates. A training session was also organized to instruct the experts on how to specify the associations between each item and its corresponding attributes, fostering a shared understanding of the relationships. Lee and Sawaki (2009b) argue that brainstorming potential attributes is particularly effective when retrofitting CDMs to existing non-diagnostic tests lacking attribute information for each test item.

Research indicates that because of the opinion-based nature of the Q-matrix development process, certain entries in the Q-matrix are prone to be inaccurately specified, as highlighted by Nájera et al. (2020). Given the subjective nature of selecting and refining attributes, empirical validation procedures were deemed essential to identify and correct any potential misalignments within the Q-matrix. In this study, the validation method proposed by de la Torre and Chiu (2016) was employed for Q-matrix validation. Notably, de la Torre and Chiu (2016) introduced a discrimination index applicable within a broad spectrum of CDMs encompassed by the G-DINA model. This index aids in empirically validating Q-matrix specifications by identifying and rectifying inaccurately specified entries in the Q-matrix while preserving correct entries. The underlying principle guiding this methodology rests on the premise that a properly specified $q$-vector should effectively discriminate between latent groups concerning the likelihood of item success. Conversely, an incorrectly specified $q$-vector would yield more homogeneous probabilities of success across the designated latent groups. Consequently, this approach offers a systematic framework for refining Q-matrix specifications based on empirical analyses conducted within the G-DINA model framework. Before executing the validation method, the plausibility of the initial Q-matrix was assessed using various absolute fit statistics (e.g., M2, SRMSR, and RMSEA2, and three residual-based statistics: transformed correlation [r], log-odds ratio [l], and proportion correct [p]). All adjustments suggested by the model were subsequently reviewed by the experts, and only those deemed theoretically plausible were implemented. These changes were made with an eye toward ensuring model fit and theoretical consistency.

### *4.4. Data Analysis*

In order to create and verify a Q-matrix for the listening section of the IELTS, a three-step procedure was undertaken. Initially, researchers formulated a preliminary Q-matrix based on existing models, and classifications of L2 listening comprehension, drawing from earlier research on the utilization of CDMs to listening comprehension, and incorporating input from the panel of experts. Subsequently, the approach developed by de la Torre and Chiu (2016) was utilized to empirically confirm the preliminary Q-matrix, employing the GDINA package version 2.9.4 (Ma et al., 2023) in R software (R Core Team, 2024). Researchers and subject matter experts meticulously reviewed all modifications suggested by the software, incorporating those in line with the theoretical framework of L2 listening into the Q-matrix while disregarding those that were not. Additionally, the analysis included the examination of mesa plots for individual items, the assessment of the heatmap plot illustrating item pair dependencies, and the scrutiny of item-level fit statistics. Finally, the fit of the G-DINA model was evaluated for the initial and revised Q-matrices utilizing various absolute fit indices to evaluate the congruence between the model and the observed data:

1- M2 (Chen & Thissen, 1997) represents the mean discrepancy between the predicted response frequencies generated by the model and the actual observed frequencies. A significant $p$-value signals

the presence of item dependency violations and inadequacies in the alignment between the model and the observed data (Hu et al., 2016);

2- $RMSEA_2$ (the root mean square error of approximation fit index for M2) is a measure of discrepancy between the observed covariance matrix and model-implied covariance matrix per degree of freedom" (Chen, 2007, p. 467). According to Maydeu-Olivares and Joe (2014), RMSEA2 values below 0.05 are indicative of a favorable fit. Additionally, Hooper et al. (2008) propose that "models with RMSEA2 values under 0.06 demonstrate an acceptable level of fit.

3- The standardized root mean squared residual (SRMSR) measures the average of standardized residuals between observed correlations and the correlations expected by the model across all pairs of items (Chen, 2007). Maydeu-Olivares (2013, p. 84) suggests that values below 0.05 suggest a negligible degree of misfit. Conversely, Hu and Bentler (1999) contend that an optimal range for SRMSR lies between 0 and 0.08.

Three item-level residual-based statistics were also scrutinized (Chen et al., 2013, p. 126): (1) Transformed correlations ($r$) represent the residual discrepancy between the Fischer-transformed correlation of item pairs predicted by the model and the observed correlation; (2) Proportion correct ($p$) measures the residual difference between the observed and predicted proportion of examinees' correct answers to a set of test items; and (3) Log-odds ratio ($l$) indicates the residual difference between the observed and predicted log-odds ratios of item pairs. Lower values suggest a better model-data fit. When the model fits well with the data, these residual-based statistics should approximate zero for all items (Chen et al., 2013). Values that are not significantly different from zero, indicated by Bonferroni adjusted $p$-values > 0.05, suggest a model with a good fit.

## 4. Results

### *4.1. Identifying Attributes*

After examining the relevant literature, and consulting with the panel of experts, six attributes were identified: (1) *Linguistic knowledge (LKA)* (Aitken, 1978; Wolfgram et al., 2016), *(2) understanding prosodic patterns (UPP)* (Aitken, 1978; Vandergrift, 2007), *(3) ability to understand and make paraphrases (PAR)* (Goh & Aryadoust, 2015; Wagner, 2004), *(4) ability to understand specific factual information such as names, numbers, and so forth (UFI)* (Sawaki et al., 2009), *(5) ability to understand explicit information (UEI)* (Field, 2009), *and (6) ability to make inference (INF)* (Tsui & Fullilove, 1998; Wagner, 2004). Finally, an initial Q-matrix was formulated, as depicted in Table 2.

**Table 2**
*The Initial Q-matrix*

| Items | LKA | UEI | UFI | PAR | UPP | INF |
|---|---|---|---|---|---|---|
| **1** | 0 | 0 | 1 | 0 | 1 | 0 |
| **2** | 0 | 0 | 1 | 0 | 1 | 0 |
| **3** | 1 | 1 | 0 | 0 | 0 | 0 |
| **4** | 0 | 0 | 1 | 0 | 1 | 0 |
| **5** | 0 | 0 | 1 | 0 | 0 | 0 |
| **6** | 1 | 1 | 0 | 0 | 1 | 0 |
| **7** | 1 | 1 | 0 | 0 | 1 | 0 |
| **8** | 0 | 0 | 1 | 0 | 1 | 0 |
| **9** | 1 | 1 | 0 | 0 | 1 | 0 |
| **10** | 1 | 0 | 1 | 0 | 1 | 0 |
| **11** | 1 | 0 | 1 | 1 | 0 | 0 |
| **12** | 1 | 0 | 1 | 1 | 0 | 0 |
| **13** | 1 | 0 | 1 | 1 | 0 | 0 |
| **14** | 1 | 0 | 1 | 1 | 0 | 0 |
| **15** | 1 | 0 | 1 | 1 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **16** | 1 | 0 | 1 | 1 | 0 | 0 |
| **17** | 1 | 0 | 1 | 1 | 0 | 0 |
| **18** | 1 | 0 | 1 | 1 | 0 | 0 |
| **19** | 1 | 0 | 1 | 1 | 0 | 0 |
| **20** | 1 | 1 | 0 | 1 | 0 | 0 |
| **21** | 1 | 1 | 0 | 0 | 1 | 0 |
| **22** | 1 | 1 | 0 | 0 | 1 | 0 |
| **23** | 1 | 1 | 0 | 0 | 1 | 0 |
| **24** | 1 | 1 | 0 | 0 | 1 | 0 |
| **25** | 1 | 1 | 0 | 1 | 1 | 0 |
| **26** | 1 | 1 | 0 | 1 | 1 | 0 |
| **27** | 1 | 1 | 0 | 1 | 1 | 0 |
| **28** | 1 | 1 | 0 | 1 | 1 | 0 |
| **29** | 1 | 0 | 0 | 1 | 1 | 1 |
| **30** | 1 | 0 | 0 | 1 | 1 | 1 |
| **31** | 1 | 0 | 1 | 1 | 0 | 1 |
| **32** | 1 | 0 | 1 | 1 | 0 | 1 |
| **33** | 1 | 1 | 0 | 1 | 0 | 0 |
| **34** | 1 | 1 | 0 | 1 | 1 | 1 |
| **35** | 1 | 1 | 0 | 1 | 0 | 0 |
| **36** | 1 | 0 | 1 | 1 | 0 | 0 |
| **37** | 1 | 1 | 0 | 1 | 0 | 1 |
| **38** | 1 | 1 | 0 | 1 | 0 | 0 |
| **39** | 1 | 0 | 1 | 1 | 1 | 0 |
| **40** | 1 | 1 | 0 | 1 | 0 | 0 |

### *4.2. Q-matrix Validation and Checking the Fit of the G-DINA Model*

As demonstrated in Table 3, The fit statistics generated by the G-DINA model were deemed unsatisfactory. The significant value of M2 indicates an inadequate model fit to the data. While the RMSEA2 and SRMSR values were lower than 0.05, indicating an acceptable fit, it is noteworthy that the upper bound confidence interval of RMSEA2 exceeded 0.05. A detailed overview of the absolute item-level fit indices of the G-DINA model is provided in Table 4. The significance level of a Z-score can be modified through the Bonferroni correction. For a significance level of $\alpha = 0.01$, a critical Z-score of 4.17 is established. Chen et al. (2013) posit that a Z-score surpassing this cut-off value suggests inadequate model fit. As demonstrated in Table 4, the G-DINA model demonstrated an adequate fit to the data according to proportion correct values (e.g., Max $Z = 0.20 < 4.17$; adjusted *p*-value $> 0.05$). However, log odds ratio and the adjusted *p*-values for the transformed correlation were significant, indicating potential misalignments in the initial Q-matrix (Sorrel et al., 2017). Thus, these findings underscore the necessity for revisions to the initial Q-matrix.

**Table 3**
*G-DINA Fit Indices for the Initial and Final Q-matrices*

| Models | Npar | $M_2$ (*p*-value) | RMSEA2 | RMSEA2. CI1 | RMSEA2. CI2 | SRMSR | -2log likelihood | AIC | BIC |
|---|---|---|---|---|---|---|---|---|---|
| Initial Q-matrix | 461 | 426.09 (0.008) | 0.035 | 0.015 | 0.008 | 0.020 | 33392 | 34314 | 36485 |
| Final Q-matrix | 499 | 358.00 (0.078) | 0.034 | 0.012 | 0 | 0.018 | 33292 | 34290 | 36640 |

*Note.* Npar = Number of parameters; CI: Confidence Intervals

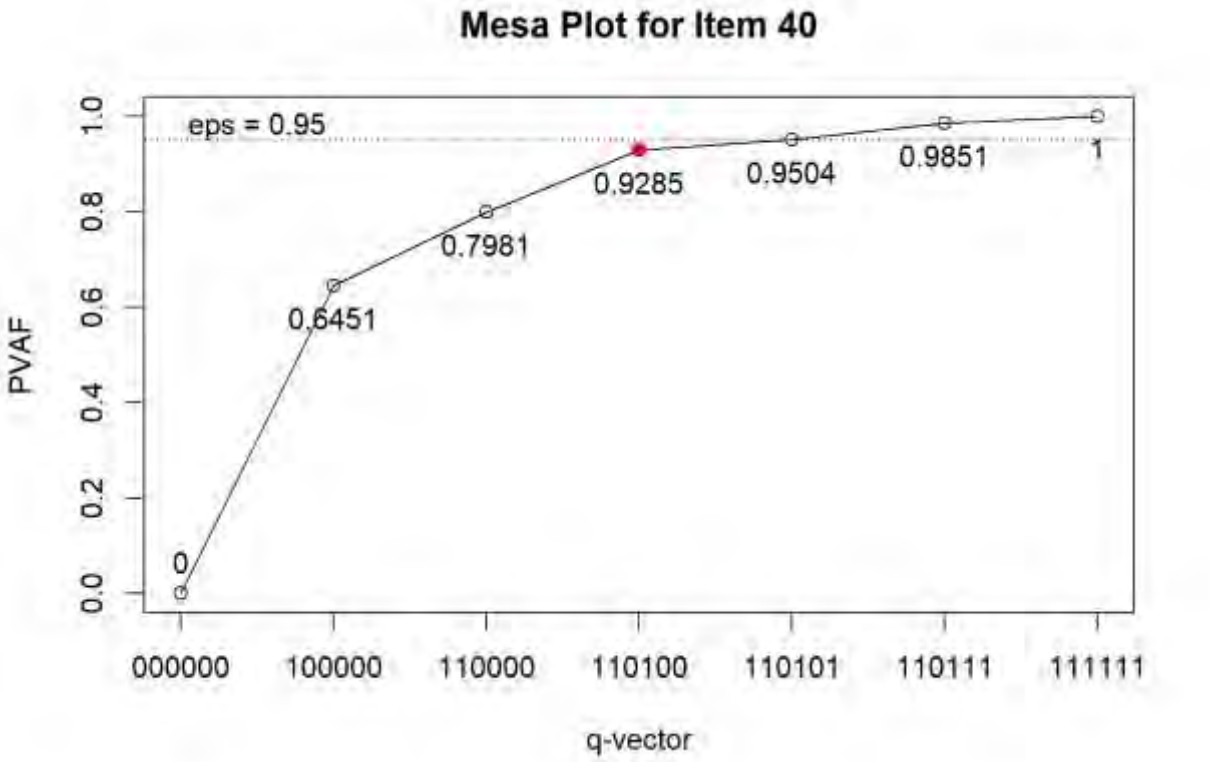
**Table 4**
*G-DINA Item Fit Indices for the Initial and Final Q-matrices*

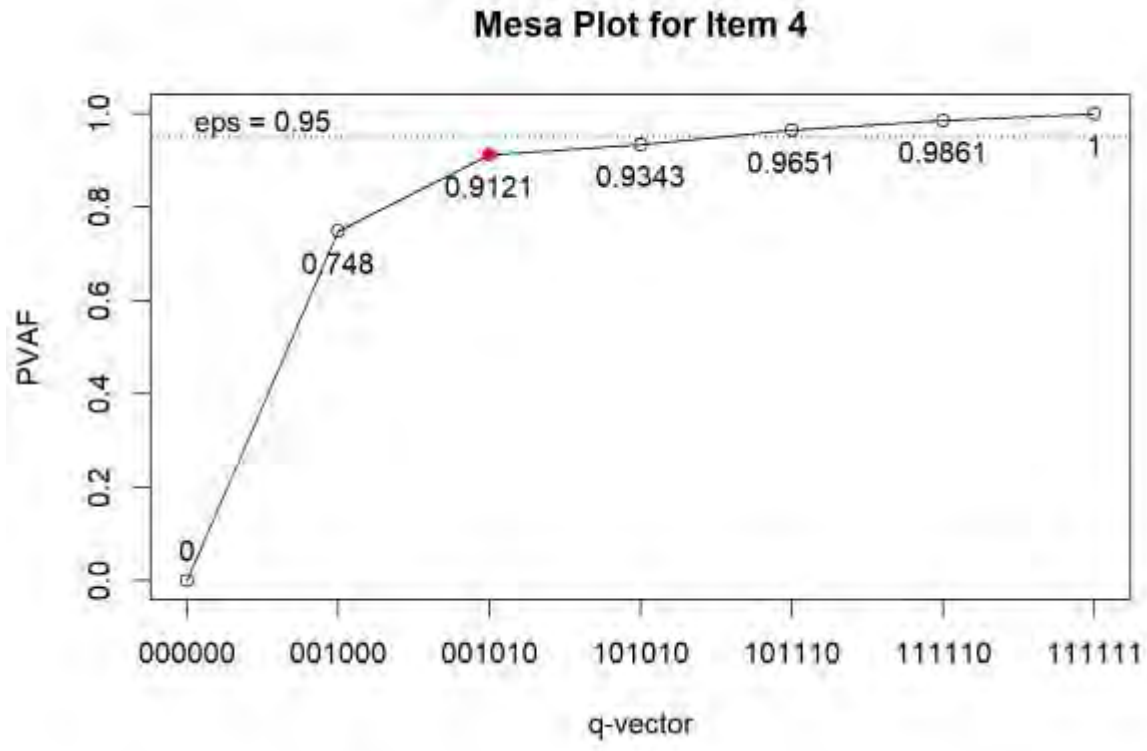
| | | mean[stats] | max[stats] | max[z.stats] | *p*-value | adj.*p*-value |
|---|---|---|---|---|---|---|
| Initial Q-matrix | Proportion correct | 0.0012 | 0.0035 | 0.20 | 0.842 | 1.0000 |
| | Transformed correlation | 0.0274 | 0.1752 | 5.01 | 0.000 | 0.0004 |
| | Log odds ratio | 0.1465 | 0.8862 | 4.47 | 0.000 | 0.0061 |
| Final Q-matrix | Proportion correct | 0.0010 | 0.0024 | 0.145 | 0.885 | 1.0000 |
| | Transformed correlation | 0.0269 | 0.1766 | 5.047 | 0.000 | 0.0004 |
| | Log odds ratio | 0.1428 | 0.7089 | 4.963 | 0.000 | 0.0005 |

Several adjustments, following de la Torre and Chiu's (2016) methodology, were proposed for the initial Q-matrix. In two instances (specifically, for Items 3 and 7), the recommendation was to change existing entries of 1s to 0s. However, there were also suggestions for introducing new entries of 1s into the Q-matrix for several items (Items 2, 3, 5, 9, 10, 11, 12, and 19). It's important to emphasize that both researchers and experts meticulously examined and evaluated each proposed modification for theoretical validity, considering only those that were theoretically sound.

To assess the effectiveness of the proposed modifications, the mesa plots for all items were examined. The mesa plot, as described by de la Torre and Ma (2016), "is a line chart and serves a function akin to the scree plot in factor analysis (Ma, 2019, p. 309). This plot illustrates *q*-vectors along the *x*-axis for varying numbers of *K* attributes, alongside their corresponding proportion of variance accounted for (PVAF) on the *y*-axis. The arrangement of *q*-vectors from lowest to highest PVAF suggests that the *q*-vector encompassing all attributes specifies the highest discrimination index." This is because, as noted by Nájera et al. (2019), "the specification of additional attributes leads to the differentiation among more latent groups, and so to a higher variability in the probabilities of success" (p. 7). Consequently, the optimal *q*-vector is the most straightforward one that explains a significant proportion of variance with the fewest attributes. Original *q*-vectors are denoted as red dots, while the optimal *q*-vector for each item is located on the edge of the mesa plot (de la Torre & Ma, 2016). A PVAF cutoff value of $\varepsilon$(EPS) = 0.95 is established. The mesa plots for Items 4 and 40 are shown in Figure 1, where the original *q*-vectors failed to reach the 0.95 threshold. The *q*-vectors [001010] and [110100] explain approximately 91% and 93% of the variance in success, respectively. These *q*-vectors are potentially the optimal choices for the items since they are situated on the edge of the mesa.

**Figure 1**
*Mesaplots for Two Items*

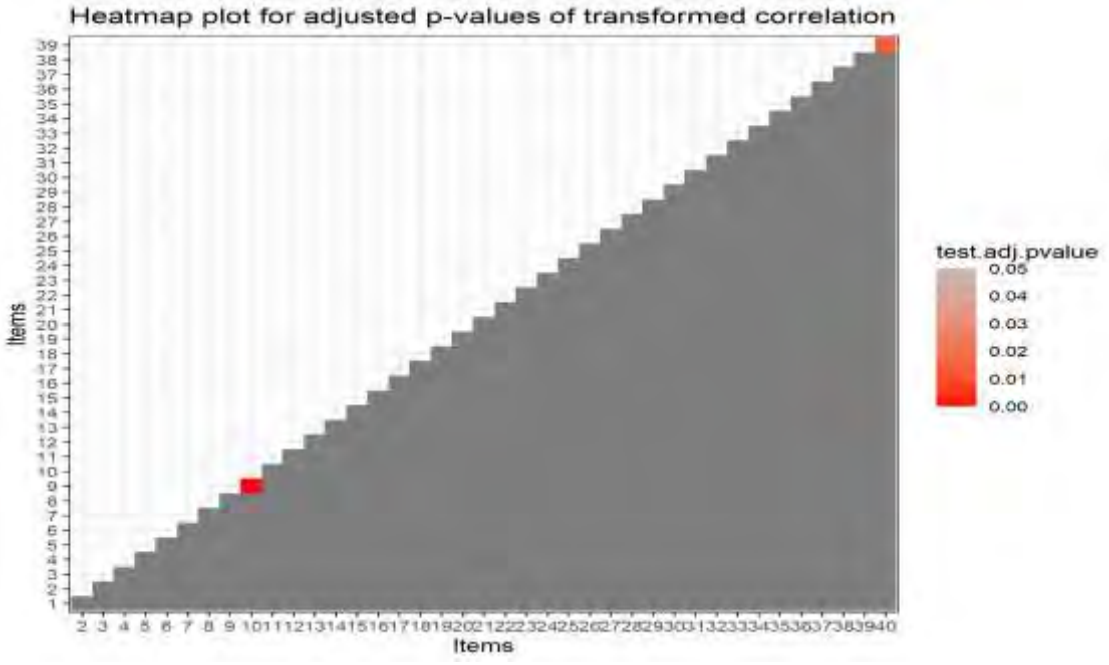


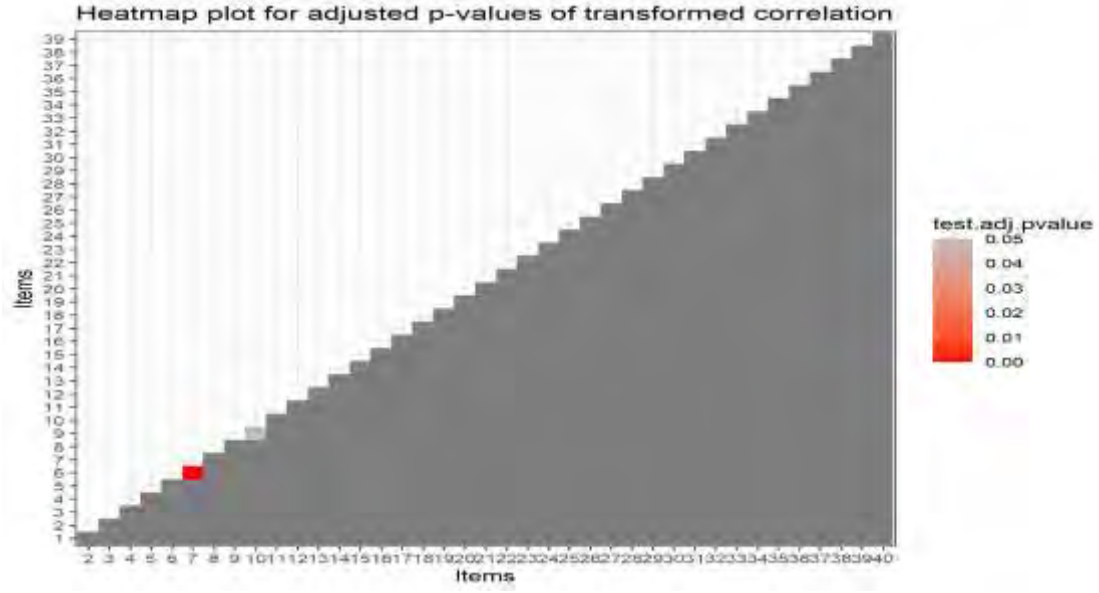
Using a heatmap plot, we also examined dependencies among pairs of items by analyzing transformed correlations. As can be seen in Figure 2, on the heatmap plots, "items are represented on the *y*- and *x*-axes, and the last and the first items are omitted on both axes. The shading area corresponds to Bonferroni adjusted *p*-values for all item pairs, where red squares indicate *p*-values lower than 0.05, signifying inadequate fit, and grey squares represent *p*-values higher than 0.05, indicating adequate fit" (Ma, 2019, p. 301). The presence of significant dependencies among certain item pairs in the initial Q-matrix is shown in Figure 2a. By applying the sensible modifications, dependencies among item pairs for the final Q-matrix were decreased. Although there is a red area on the plot, the G-DINA model fit showed an improvement of the model fit. Fit statistics for the initial and final Q-matrices are illustrated in Tables 3 and 4.

**Figure 2**

*Heatmap Visualization of Adjusted p-values for the Initial and Final Q-matrices*




**(a)**




**(b)**

The relative fit indices values, such as -2log likelihood, AIC, and BIC, indicated the better performance of the final Q-matrix relative to the initial one. The non-significant *p*-value (e.g., 0.25) suggests that the model has a good fit to the data. Concerning RMSEA2, both the value of the G-DINA model (e.g., 0.0122) and its lower and upper bounds were below 0.05. Additionally, the SRMSR value was below 0.05. Finally, it must be noted that although the values of absolute fit statistics at the test-level showed a sufficient model-date fit, the log odds ratio and adjusted *p*-values for transformed correlation remain significant. The final Q-matrix is presented in Table 5.

**Table 5**
*The Final Q-matrix*

| Items | LKA | UEI | UFI | PAR | UPP | INF |
|---|---|---|---|---|---|---|
| **1** | 0 | 0 | 1 | 0 | 1 | 0 |
| **2** | 0 | 1* | 1 | 0 | 1 | 0 |
| **3** | 1 | 0 | 1* | 0 | 0 | 0 |
| **4** | 0 | 0 | 1 | 0 | 1 | 0 |
| **5** | 1* | 0 | 1 | 0 | 1* | 0 |
| **6** | 1 | 1 | 0 | 0 | 1 | 0 |
| **7** | 0 | 1 | 0 | 0 | 1 | 0 |
| **8** | 0 | 0 | 1 | 0 | 1 | 0 |
| **9** | 1 | 1 | 1* | 0 | 1 | 0 |
| **10** | 1 | 0 | 1 | 1* | 1 | 0 |
| **11** | 1 | 0 | 1 | 1 | 1* | 0 |
| **12** | 1 | 0 | 1 | 1 | 1* | 0 |
| **13** | 1 | 0 | 1 | 1 | 0 | 0 |
| **14** | 1 | 0 | 1 | 1 | 0 | 0 |
| **15** | 1 | 0 | 1 | 1 | 0 | 0 |
| **16** | 1 | 0 | 1 | 1 | 0 | 0 |
| **17** | 1 | 0 | 1 | 1 | 0 | 0 |
| **18** | 1 | 0 | 1 | 1 | 0 | 0 |
| **19** | 1 | 0 | 1 | 1 | 1* | 0 |
| **20** | 1 | 1 | 0 | 1 | 0 | 0 |
| **21** | 1 | 1 | 0 | 0 | 1 | 0 |
| **22** | 1 | 1 | 0 | 0 | 1 | 0 |
| **23** | 1 | 1 | 0 | 0 | 1 | 0 |
| **24** | 1 | 1 | 0 | 0 | 1 | 0 |
| **25** | 1 | 1 | 0 | 1 | 1 | 0 |
| **26** | 1 | 1 | 0 | 1 | 1 | 0 |
| **27** | 1 | 1 | 0 | 1 | 1 | 0 |
| **28** | 1 | 1 | 0 | 1 | 1 | 0 |
| **29** | 1 | 0 | 0 | 1 | 1 | 1 |
| **30** | 1 | 0 | 0 | 1 | 1 | 1 |
| **31** | 1 | 0 | 1 | 1 | 0 | 1 |
| **32** | 1 | 0 | 1 | 1 | 0 | 1 |
| **33** | 1 | 1 | 0 | 1 | 0 | 0 |
| **34** | 1 | 1 | 0 | 1 | 1 | 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **35** | 1 | 1 | 0 | 1 | 0 | 0 |
| **36** | 1 | 0 | 1 | 1 | 0 | 0 |
| **37** | 1 | 1 | 0 | 1 | 0 | 1 |
| **38** | 1 | 1 | 0 | 1 | 0 | 0 |
| **39** | 1 | 0 | 1 | 1 | 0 | 0 |
| **40** | 1 | 1 | 0 | 1 | 0 | 0 |

*Note.* Stared items show that these items were 0 in the initial Q-matrix, and finally changed into 1./ Underlined items show that these items were 1 in the initial Q-matrix, and finally changes into 0.

## 5. Discussion

The present study aimed to develop and validate a Q-matrix tailored specifically for the listening section of the IELTS, a high-stakes examination. To validate the Q-matrix, its fit was scrutinized using the G-DINA model, both at the test and item levels. Through the application of the CDM to the test, misspecifications were identified, and subsequent modifications were suggested by the software. However, not all proposed modifications were deemed feasible or appropriate for implementation. An in-depth examination of mesa and heatmap plots, alongside the proposed modifications, prompted further deliberation with the expert panel. Collaboratively, a subset of the suggested modifications was incorporated into the Q-matrix, culminating in the derivation of a final iteration that exhibited favorable alignment with the G-DINA model. The final Q-matrix is valid and sensible.

However, a comparative analysis between the pre- and post-modification versions of the Q-matrix revealed minimal alterations. This phenomenon could be attributed to the interdependence of items within each section of the IELTS test. Notably, the IELTS consists of four distinct sections, with each section comprising 10 questions pertaining to a specific thematic domain.

Upon further examination of the results, it became evident that although the G-DINA model exhibited a satisfactory fit with the final Q-matrix at the test level, it lacked significant alignment at the item-level. A potential justification for this discrepancy could be the observed dependency between item 6 and item 7, as depicted in the heatmap plot (b). This inter-item dependency may stem from the nature of certain items within the listening sections of the IELTS, wherein some items share identical stems. Alternatively, this dependency could be attributed to the rapid succession of stimuli in the audio input of the test. Analysis of the test content revealed that the responses to items 6 and 7 were presented in the audio file in close proximity to each other, allowing minimal time for test takers to transition from answering one question to the next. Consequently, individuals may find themselves preoccupied with formulating a response to item 6, thereby impeding their ability to effectively process and respond to item 7 within the allotted timeframe.

This phenomenon underscores the complexity of listening tests, wherein test takers are tasked with simultaneously listening, reading, and writing. The observed issue appears to function more as a distraction rather than a valid assessment of listening comprehension. This observation aligns with the discovery by Coleman and Heap (1998) that the consecutive presentation of two questions in the IELTS listening test could potentially lead to comprehension difficulties for the test takers. Such items may inadvertently assess factors such as reading speed and memory span, which are unrelated to the latent trait of listening comprehension. According to Shohamy and Inbar (1991), items on listening comprehension tests that require test takers to focus on memory skills and trivial details impose a significant burden on the test takers' memory load and are not conducive to effective assessment.

The simultaneous exposure to oral and written stimuli in While-Listening Performance (WLP) tests can impede note-taking abilities. Consequently, individuals who lag behind in processing the flow of written and oral information may overlook certain items. Previous research has indicated that this oversight is not necessarily indicative of deficient listening abilities, but could instead stem from potential constraints such as reading proficiency, memory span (Hildyard & Olson, 1978), test-taking strategies (Bachman, 1990), test wiseness (Bachman, 1990; Kunnan, 1995), or other limiting factors (Field, 2009; Meng & Fu, 2023). As such, the outcomes of this study underscore the need for meticulous

item design to ensure that each item accurately assesses the intended construct without introducing extraneous factors that may confound test takers.

Upon analyzing the test content and delineating the associated attributes, it was noted that none of the attributes directly pertained to functional knowledge. This involves a test taker's ability to discern the function or illocutionary force of a statement or text, interpreting the intended meaning within context. This aligns with Weir's (2005) critique that While Listening Performance Tests, including the IELTS, inadequately assess pragmatic knowledge. Pragmatic knowledge, a higher-order skill, requires interpretation and inference-making beyond literal comprehension and paraphrasing (Hildyard & Olson, 1978).

Given that all research encounters limitations, this study also acknowledges potential constraints that may affect its outcomes. Firstly, while the Q-matrix developed in this study underwent thorough empirical validation, it is essential to recognize that it represents just one possible configuration for the listening section of the IELTS exam. Subsequent studies might explore alternative approaches to determine the optimal granularity of attributes and employ diverse strategies for Q-matrix construction. Nevertheless, it's important to note that although incorporating numerous skills can enhance diagnostic information, it can strain the statistical modeling capacity within the test's length. Test developers must balance the number of attributes assessed with the test's overall length, considering theoretical, technical, and practical factors as advised by Jang (2009).

Secondly, employing a CDA approach in this study involved retrofitting a non-diagnostic test, raising concerns about the credibility of conclusions drawn regarding test takers' skill mastery profiles. Unlike true CDMs, retrofitting existing tests necessitates careful test and attribute specifications, potentially compromising the quality of CDA in offering detailed diagnostic information. Nonetheless, the process of retrofitting can play a vital role in advancing cognitive diagnostic assessment (Lee and Sawaki, 2009a; Yumsek, 2023), by assessing the feasibility of extracting valuable cognitive diagnostic information from existing assessments before embarking on the resource-intensive process of designing new diagnostic tests, as advocated by Ravand and Baghaei (2019).

Additionally, although the sample size of this study (N=820) holds practical significance, it may be deemed inadequate for CDM application (de la Torre & Lee, 2013;
Ma et al., 2016). However, gathering extensive data in educational settings poses considerable challenges due to constraints in data collection procedures. Limited research on the impact of sample size on CDM application indicates potential effects on parameter recovery and fit indices. Notably, some researchers suggest small sample sizes can be promising for identifying appropriate CDMs. Cognitive diagnostic assessment offers rich diagnostic insights into students' learning status, highlighting the need for educational assessments grounded in a CDM framework, which requires interdisciplinary collaboration.

Furthermore, in the context of second language acquisition, age plays a vital role in language learning and proficiency development, significantly contributing to individual differences (Muñoz, 2006; Singleton & Ryan, 2004). However, this study did not focus on how age might influence listening proficiency or test performance. We hope that future researchers will address this issue in their work. Further research on this topic is recommended.

Besides, the expert panel involved in developing the Q-matrix consisted entirely of non-native English speakers, which could potentially introduce biases or limitations in understanding the nuances of English language proficiency. However, the researchers did not have access to native English-speaking experts. If this research were conducted in a different country with native speaker experts, the results might vary, given the subjective nature of attribute extraction and validation.

Finally, it is imperative to recognize the study's scope limitations, as the test participants did not represent a diverse, global population. Even, it was not feasible to collect data from all regions of Iran. Consequently, participants were selected from some universities and English language institutes. This selection may not accurately reflect the diversity of EFL learners in Iran, potentially introducing sampling bias and limiting the generalizability of the findings to other contexts or populations. Future research efforts should aim to diversify their participant pool to enhance the generalizability of findings across varied demographics and contexts.

## 6. Conclusion and Implications

The outcomes of this study hold implications across three significant dimensions. Firstly, from a theoretical standpoint, the findings serve to elucidate the underlying attributes of the listening test. Consequently, these results can inform the refinement of existing models, contributing to the evolution of theoretical frameworks in this domain.

Secondly, at a methodological level, this study extends prior research efforts by furnishing a Q-matrix that is both valid and coherent. This advancement enhances the methodological rigor of cognitive diagnostic modeling in the context of listening comprehension assessments. Moreover, since the model used to develop the Q-matrix in this study is a general one, other researchers can utilize the designed Q-matrix for both compensatory and non-compensatory models. This flexibility allows for the revision and refinement of attributes in future studies.

Thirdly, from a practical perspective, two key points merit attention. In the realm of test preparation, this study sheds light on the interdependencies among test questions, underscoring the need for meticulous design considerations in high-stakes assessments like the IELTS. Specifically, it advocates for thoughtful input design, emphasizing the importance of spacing between answers to afford test takers sufficient time for response formulation without the burden of simultaneous processing. These attributes can also enhance the design of tests tailored to cognitive structures. Additionally, educators and institutions responsible for preparing students for such assessments should pay attention to the attributes delineated in this study, as well as those identified in related research. Tailoring instructional content to align with these attributes and associated constructs can enhance students' readiness for high-stakes tests, thereby facilitating more effective test preparation strategies (Helm et al, 2022).

Taking into account all these considerations and the potential applications of implementing the results of this study, it's important to acknowledge that certain challenges may arise during the implementation of the findings. Since test takers were not directly involved in defining the attributes, there is a possibility that the attributes identified by content experts may not fully align with the perceptions and attributes held by test takers as they listen, comprehend, and respond to test items. It might be beneficial to incorporate a think-aloud protocol and other empirical methods to identify and validate cognitive attributes more effectively.

**Declaration of Conflicting Interests**

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

**Funding**

The authors received no specific funding for this work from any funding agencies.

**References**

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2005). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement, Issues, and Practice, 22* (3), 37–53. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

Aitken, K. G. (1978). *Measuring listening comprehension in English as a second language* [Research Report]. TEAL Occasional Papers, Volume 2. British Columbia Association of Teachers of English as an Additional Language, Vancouver. https://eric.ed.gov/?id=ED155945

Alavi, S. M., Kaivanpanah, Sh., & Panahi Masjedlou, A. (2018). Validity of the listening module of international English language testing system: Multiple sources of evidence. *Language Testing in Asia, 8*(8), 1–17. https://doi.org/10.1186/s40468-018-0057-4

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An Individual differences approach. *Language Learning*, *62*, 49–78. https://doi.org/10.1111/j.1467-9922.2012.00706.x

Aryadoust, V. (2011). Constructing validity arguments for the speaking and listening modules of international English language testing system. *The Asian ESP Journal, 7*(2), 28–54. https://asian-esp-journal.com/wp-content/uploads/2016/01/AESP-Volume7-Issue2-April-2011.pdf

Aryadoust, V. (2012). Differential Item Functioning in While-Listening Performance Tests: The Case of International English Language Testing System (IELTS) Listening Module. *International Journal of Listening, 26*(1), 40–60. https://doi.org/10.1080/10904018.2012.639649

Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency.* Cambridge Scholars Publishing.

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 RRUM. *International Journal of Listening*, *35*(1), 29–52. https://doi.org/10.1080/10904018.2018.1500915

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Badger, R, & Yan, X. (2009). The use of tactics and strategies by Chinese students in the listening component of IELTS. *IELTS Research Report*, 9, 67–96. British Council and IELTS Australia. https://ielts.org/researchers/our-research/research-reports/the-use-of-tactics-and-strategies-by-chinese-students-in-the-listening-component-of-ielts

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (Report No. RM-00-07) [Research Report]. TOEFL Monograph Series No. MS-19. Educational Testing Service. https://www.ets.org/research/policy_research_reports/publications/report/2000/iciu.html

Birenbaum, M., Kelly, A. E., & Tatsuoka, K. K. (1993). Diagnosing knowledge states in Algebra using the Rule-space model. *Journal for Research in Mathematics Education, 24*(5), 442–459. https://doi.org/10.2307/749153

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

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

Brown, H. D. (2007). *Teaching by principles: An interactive approach to language pedagogy*. Longman.

Buck, G. (2001). *Assessing listening.* Cambridge University Press. https://doi.org/10.1017/CBO9780511732959

Buck, G., & Tatsuoka, L. (1998). Application of the rule-space procedure to language testing examining attributes of a free response listening test. *Language Testing, 15*, 119–157. https://doi.org/10.1177/026553229801500201

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chen, H., Cai, Y., & de la Torre, J. (2022). Investigating second language (L2) reading subskill associations: A cognitive diagnosis approach. *Language Assessment Quarterly*, *20*(2), 166–189. https://doi.org/10.1080/15434303.2022.2140050

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, J, de la Torre, J, & Zhang, Z. (2013). Relative and absolute fir evaluation in cognitive diagnostic modelling. *Journal of Educational Measurement, 50*(2), 123–140. https://doi.org/10.1111/j.1745-3984.2012.00185.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289. https://doi.org/10.2307/1165285

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*(8), 598–618. https://doi.org/10.1177/01466216134884

Coleman, G., & Heap, S. (1998). The misinterpretation of directions for the questions in the Academic Reading and Listening sub-tests of the IELTS test (Research Report No. 1) *IELTS Research Report*, British Council and IELTS Australia. https://ielts.org/researchers/our-research/research-reports/the-misinterpretation-of-directions-for-the-questions-in-the-academic-reading-and-listening-sub-tests-of-the-ielts-test

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. https://doi.org/10.1007/BF02310555

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*(3), 163–183. https://doi.org/10.1177/0146621608320523

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

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

de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*, 355–373. https://doi.org/10.1111/jedm.12022

de la Torre, J., & Ma, W. (2016, August). *Cognitive diagnosis modeling: A general framework approach and its implementation in R.* A Short Course at the Fourth Conference on Statistical Methods in Psychometrics, Columbia University, New York.

de la Torre, J., Qiu, X. L., & Santos, K. C. (2022). An empirical Q-matrix validation method for the polytomous GDINA model. *Psychometrika*, *87*(2), 693–724. https://doi.org/10.1007/s11336-021-09821-x

Dong, Y., Ma, X. Wang, Ch., & Gao, X. (2021). An Optimal Choice of Cognitive Diagnostic Model for Second Language Listening Comprehension Test. *Frontiers in Psychology, 12,* 1–12. https://doi.org/10.3389/fpsyg.2021.608320

Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *Modern Language Journal, 77*, 180–191. https://doi.org/10.1111/j.1540-4781.1993.tb01962.x

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. https://www.ijlt.ir/article_114295.html

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. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179–197. https://doi.org/10.1037/0033-2909.93.1.179

Field, J. (2009). *Listening in the language classroom*. Cambridge University Press. https://doi.org/10.1017/CBO9780511575945

Field, J. (2012). The Cognitive Validity of the Lecture-based question in the IELTS Listening Paper. In: *IELTS Collected Papers 2: Research in reading and listening assessment*. Cambridge University Press. http://hdl.handle.net/10547/225496

Flowerdew, J., & Miller, L. (2005). *Second language listening: Theory and practice*. Cambridge University Press. https://doi.org/10.1017/CBO9780511667244

Gao, L., & Rodgers, T. (2007, April). Cognitive-psychometric modeling of the MELAB reading items. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Goh, C. C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*, *28*(1), 55–75. https://doi.org/10.1016/S0346-251X(99)00060-3

Goh, C. & Aryadoust, V. (2015). Examining the notion of listening subskill divisibility and its implications for second language listening. *International Journal of Listening, 29*(3), 109-133. https://doi.org/10.1080/10904018.2014.936119

Harding, L, Alderson, JC, & 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.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301–321. https://doi.org/10.1111/j.1745-3984.1989.tb00336.x

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

Helm, C., Warwas, J. & Schirmer, H. (2022). Cognitive diagnosis models of students' skill profiles as a basis for adaptive teaching: an example from introductory accounting classes. *Empirical Research in Vocational Education and Training, 14*(9), 1–30. https://doi.org/10.1186/s40461-022-00137-3

Henson, R. A., Templin, J. L., & Willse, J. T. (2008). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191–210. https://doi.org/10.1007/s11336-008-9089-5

Hildyard, A., & Olson, D. (1978). Memory and inference in the comprehension of oral and written discourse. *Discourse Processes, 1*, 91–107. https://doi.org/10.1080/01638537809544431

Hooper, D., Coughlan, J., & Mullen, M. (2008, June 19-20). Evaluating model fit: A synthesis of the structural equation modelling literature. In A. Brown (Ed.) *Proceedings of 7th European Conference on Research Methodology for Business and Management Studies*, London, UK, 195–200.
https://books.google.com/books?hl=en&lr=&id=ZZoHBAAAQBAJ&oi=fnd&pg=PA195&ots=gXXQZtUr75&sig=10QFO9N8ecvOE4iBUi-tNFaBgwY#v=onepage&q&f=false

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to Language assessment. *Language Testing, 26*(1), 031–073. https://doi.org/10.1177/0265532208097336

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

Kang, C., Yang, Y., & Zeng, P. (2019). Q-matrix refinement based on item fit statistic RMSEA. *Applied Psychological Measurement*, *43*(7), 527–542. https://doi.org/10.1177/0146621618813104

Kim, A. Y. (2014). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2), 227–258. https://doi.org/10.1177/0265532214558457

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge University Press.

Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural modeling approach.* Cambridge University Press.

Lee, Y.-S., de la Torre, J., & Park, Y. S. (2012). Relationships between cognitive diagnosis, CTT, and IRT indices: An empirical investigation. *Asia Pacific Education Review, 13*(2), 333–345. https://doi.org/10.1007/s12564-011-9196-3

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

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

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*(3), 205–237. https://doi.org/10.1111/j.1745-3984.2004.tb01163.x

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

Li, H., & Suen, H. K. (2013). Detecting native language group differences at the subskills level of reading: A differential skill functioning approach. *Language Testing, 30*(2), 273–298. https://doi.org/10.1177/0265532212459031

Liao, Y. (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. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 5*, 37–78. University of Michigan. https://michiganassessment.org/wp-content/uploads/2020/02/20.02.pdf.Res_.InvestigatingtheConstructValidityoftheGrammarandVocabularySectionandtheListeningSectionoftheECCE-LexicoGrammaticalAbilityasaPredictorofL2ListeningAbility.pdf

Ma, W. (2019). Cognitive diagnosis modeling using the GDINA R package. In M. von Davier & Y. S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 593–601). Springer Nature. https://doi.org/10.1007/978-3-030-05584-4_29

Ma, W., & de la Torre, J. (2020). An empirical Q-matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology, 73*(1), 142–163. https://doi.org/10.1111/bmsp.12156

Ma, W., de la Torre, J., Sorrel, M., & Jiang, Zh. (2023). *GDINA: The generalized DINA model framework.* R package version 2.9.4. https://CRAN.R-project.org/package=GDINA

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement, 40*, 200–217. https://doi.org/10.1177/0146621615621717

Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement, 75*(3), 491–511. https://doi.org/10.1177/0013164414539162

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212. https://doi.org/10.1007/BF02294535

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives, 11*(3), 71–101. https://doi.org/10.1080/15366367.2013.831680

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

Meng, Y. (2013). *Developing a model of cognitive diagnostic assessment for college EFL listening* [Unpublished doctoral dissertation]. Shanghai International Studies University.

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

Min, S., & He, L. (2021). Developing individualized feedback for listening assessment: Combining standard setting and cognitive diagnostic assessment approaches. *Language Testing*, *39*(1), 1–27. https://doi.org/10.1177/0265532221995475

Muñoz, C. (2006). *Age and the rate of foreign language learning*. Multilingual Matters. https://doi.org/10.21832/9781853598937

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

Nakatsuhara, F., Inoue, C., & Taylor, L. (2017). An investigation into double-marking methods: comparing live, audio and video rating of performance on the IELTS Speaking Test. *IELTS Research Reports Online Series, 1*, 1–49. British Council and IELTS Australia. https://ielts.org/researchers/our-research/research-reports/an-investigation-into-double-marking-methods-comparing-live-audio-and-video-rating-of-performance-on-the-ielts-speaking-test

Phakiti, A. (2016). Test-takers' performance appraisals, appraisal calibration, state-trait strategy use, and state-trait IELTS listening difficulty in a simulated IELTS Listening test. *IELTS Research Reports Series, 6,* 1–140. https://s3.eu-west-2.amazonaws.com/ielts-web-static/production/Research/test-takers-performance-appraisals-appraisal-calibration-state-trait-strategy-use-and-state-trait-ielts-listening-difficulty-phakiti-2016.pdf

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

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

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

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

R Core Team (2024). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.

Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly, 17*(2), 219–240. https://doi.org/10.2307/3586651

Rost, M. (2016). *Teaching and researching: Listening* (3rd ed.). Longman.

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

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

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

Shin, S. (2008). Examining the construct validity of a web-based academic listening test: An investigation of the effects of response formats. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *6*, 95–129. University of Michigan. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=dX23NVoAAAAJ&citation_for_view=dX23NVoAAAAJ:d1gkVwhDpl0C

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question type. *Language Testing, 8*(1), 23–40. https://doi.org/10.1177/026553229100800103

Singleton, D., & Ryan, L. (2004). *Language acquisition: The age factor* (2nd ed.). Multilingual Matters. https://doi.org/10.21832/9781853597596

Snowling, M. J., & Hulme, C. (Eds.). (2005). *The science of reading: A handbook.* Blackwell.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the Rule Space Method*. Routledge. https://doi.org/10.4324/9780203883372

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Taylor, L. & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalizing the test construct. *Journal of English for Academic Purposes, 10*(2), 89–101. https://doi.org/10.1016/j.jeap.2011.03.002

Tsui, A. B. M., & Fullilove, J. (1998). Bottom-up or top-down processing as a discriminator of L2 listening performance. *Applied Linguistics*, *19*, 432–451. https://doi.org/10.1093/applin/19.4.432

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching, 40*, 191–210. https://doi.org/10.1017/S0261444807004338

Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. Routledge.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287–307. https://doi.org/10.1348/000711007x193957

Wagner, E. (2004). A construct validation study of the extended listening sections of the ECPE and MELAB. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *2*, 1–23. University of Michigan. https://michiganassessment.org/wp-content/uploads/2020/02/20.02.pdf.Res_.AConstructValidationStudyoftheListeningSectionsoftheECPEandMELAB.pdf

Wang, W., Song, L., Ding, S., Meng, Y., Cao, C., & Jie, Y. (2018). An EM-based method for Q-matrix validation. *Applied Psychological Measurement*, *42*(6), 446–459. https://doi.org/10.1177/0146621617752991

Weir, C. J. (2005). *Language testing and validation: An evidenced-based approach*. Palgrave Macmillan.

Winke, P, & Lim, H. (2014). The effects of testwiseness and test-taking anxiety on L2 listening test performance: a visual (eye-tracking) and attentional investigation. *IELTS Research Reports Series, 3*, 1–30. https://s3.eu-west-2.amazonaws.com/ielts-web-static/production/Research/effects-of-testwiseness-and-test-taking-anxiety-on-l2-listening-test-performance-winkle-et-al-2014.pdf

Wolfgramm, C., Suter, N., & Göksel, E. (2016). Examining the role of concentration, vocabulary and self-concept in listening and reading comprehension. *International Journal of Listening, 30*, 25–46. https://doi.org/10.1080/10904018.2015.1065746

Xie, Q. (2016). Diagnosing university students' academic writing in English: Is cognitive diagnostic modelling the way forward? *Educational Psychology, 37*(1), 26–47. https://doi.org/10.1080/01443410.2016.1202900

Yeldham, M. (2016). Second language listening instruction: Comparing a strategies-based approach with an interactive, strategies/bottom-up skills approach. *TESOL Quarterly, 50*(2), 394–420. https://doi.org/10.1002/tesq.233

Yi, Y. (2017). Probing the relative importance of different attributes in L2 reading and listening comprehension items: An application of cognitive diagnostic models. *Language Testing*, *34*(3), 337–355. https://doi.org/10.1177/0265532216646141

Yumsek, M. (2023). Educational L2 constructs and diagnostic measurement. *Language Testing in Asia, 13*(3), 1–23. https://doi.org/10.1186/s40468-022-00214-0