Heidari Vincheh *et al.*
Language Testing in Asia (2024) 14:42
https://doi.org/10.1186/s40468-024-00311-2

Language Testing in Asia

RESEARCH Open Access

A cognitive diagnostic approach to IELTS speaking test: unveiling the subskills and test-takers' perceptions

Maryam Heidari Vincheh[1], Azizullah Mirzaei[1*] and Ali Roohani[1]

*Correspondence:
mirzaei-a@lit.sku.ac.ir

[1] English Department, Faculty of Letters & Humanities, Shahrekord University, Shahrekord, Iran

Abstract

Although the IELTS speaking test assesses English verbal-communication competence accurately and reliably, its traditional, retroactive scoring framework stops short of providing detailed diagnostic information about individual test-takers' specific domains of proficiency or deficiency. This study employed a cognitive diagnostic assessment (CDA) framework, grounded in cognitive psychology and diagnostic, formative feedback theories, to identify the specific cognitive (sub)skills and attributes required to successfully perform on the IELTS speaking test and examine the test-takers' perceptions regarding the CDA and IELTS-speaking-assessment interface. The purpose was to integrate CDA and feedback literacy into the IELTS assessment. Adopting a cognitive/psychometric perspective, we conceptualized language proficiency as encompassing separate cognitive subskills that can be systematically measured and diagnosed. To this end, we analyzed speaking data from 500 pre-IELTS candidates and conducted 20 expert interviews. The thematic analysis employed through the MAXQDA software helped identify five specific subskills related to the IELTS speaking test. Using a similar thematic analysis procedure, we extracted five perceptive themes regarding an integrated IELTS-CDA framework from the interview data of 151 participants. The results have important implications for the practicality and efficacy of employing CDA to improve IELTS speaking interpretation and feedback.

Keywords: IELTS speaking, CDA, Feedback, Speaking subskills, Perceptions

Introduction

It is widely accepted that good speaking skills are essential to second language (L2) communicative competence, defined as the ability to use language effectively in social, academic, and professional settings (Bachman, 1990; Celce-Murcia, 2007; Hymes, 1972). Oral proficiency in languages like English improves international relations, multicultural understanding, economic development, and research collaboration (Crystal, 2012). In addition, L2 speaking allows people to go across borders for better occupation, education, and well-being (Kurata, 2011). Given the importance of L2 speaking in human contact, using valid and reliable tests or assessment procedures is essential. However, assessing a complicated, context-dependent, and variable construct, such as L2 speaking,

is always difficult (Luoma, 2004). Current assessment theory and research emphasize the use of integrated performance tests (e.g., Taylor, 2011) or dynamic assessment procedures (e.g., Derakhshan & Kordjazi, 2015; Minakova, 2020), which require careful implementation by skilled examiners. In essence, the assessment of L2 oral communication skills entails striking a balance between authenticity, validity, reliability, and practicality in light of the purported test use.

The International English Language Testing System (IELTS) speaking module is widely recognized and extensively studied as a renowned assessment tool for L2 speaking skills. The test assesses oral communication skills from basic to advanced levels by taking different in-person interviews, or one-on-one speech samples, that are rated by skilled examiners. IELTS test results are essentially used for, *inter alia*, education, immigration, and employment purposes in the USA, UK, Canada, Australia, and around the globe (Taylor & Jones, 2006). Validity arguments and reliability data regarding the IELTS-speaking module assessment (e.g., Allen, 2016; O'Loughlin, 2001; Weir & O'Sullivan, 2011) are always important, given the high-stakes decision-making purposes the test is used for. However, merging analytically evaluated components into a holistic band score does not reveal individual strengths and weaknesses (Deygers et al., 2018). The opaque score makes developmental feedback and goal-setting difficult without competency analysis (Galaczi et al., 2011; Nakatsuhara et al., 2017; Wu et al., 2021). For placement and admissions, the IELTS band score measures general oral proficiency, not competency patterns. Therefore, two people with equal test scores can have different strengths and weaknesses. Moreover, the rapid transition to online assessment for a high-stakes test such as IELTS has presented distinct problems pertaining to the secure delivery and the employed evaluation frameworks (Clark et al., 2022).

Recent research has increasingly emphasized the use of cognitive diagnostic assessment (CDA) principles to allow for or boost dimensionality and diagnostic capacities (Jang, 2005; Nakatsuhara et al., 2017). CDA, which goes beyond scores to include (metacognitive) knowledge or skill information, is rooted in cognitive psychology, information processing, as well as formative (self-)assessment, and draws upon feedback theory. These views suggest that mastering complex fields requires the development of new skills, methods, knowledge, and diagnostic feedback loops (Rupp & Templin, 2008). CDA provides a more complete learner profile by identifying and enacting these traits (Leighton & Gierl, 2007). CDA improves speech assessment multidimensionality, promotes diagnostic values, and activates feedback processes while retaining validity and reliability (Sawaki et al., 2009).

In IELTS, as noted, candidates receive an overall score on the speaking test based on pre-set criteria. Test-takers do not receive a complete test score breakdown for each criterion. The ambiguity prevents applicants from assessing their speech-domain strengths and limitations. To bridge this gap, CDA seeks to assess a test-taker's cognitive abilities and metacognitive processes. Although CDA in high-stakes language tests presents challenges, it has the potential to offer test-takers more detailed and focused feedback. Research on modeling diagnostic cognition and formative feedback through assessment is needed to address this lacuna in standardized large-scale language-testing approaches and procedures in a globalized world (Hamid et al., 2019), in general, and the gap between the IELTS speaking assessment and the CDA, in particular. This

absence of diagnostic transparency in assessing speaking adversely affects test-takers' interpretation of the band scores and hinders the implementation of focused assessment and feedback for sustained improvement. Although the advent of CDA in the context of high-stakes language assessment shows promise, there are still unresolved basic challenges and issues that need to be addressed.

Apart from this perceived need to bridge the gap between assessing L2 speaking and CDA, the IELTS exam has emerged as a crucial gatekeeping proficiency measure for Iranian students aspiring to get admission to higher-education programs, both domestically and internationally, thereby pursuing occupation and emigration dreams against the nation's evolving sociopolitical backdrop. Many of these students perceive achieving high scores on the IELTS exam as a vital achievement that can unlock educational and professional prospects for them that might otherwise be challenging to attain. Iranian IELTS test-takers and institutes are increasingly seeking more thorough formative feedback or diagnostic inferences about learners' strengths and weaknesses in L2 skills of interest for preparatory, learning, and analytical purposes. Traditional IELTS holistic scoring methods provide an overall proficiency score but fail to describe this much-needed information. Given the perceived lack of CDA studies on assessment frameworks for L2 speaking as well as the high-stakes testing situation of English for Iranian L2 learners, this study aimed to, firstly, tap into the specific cognitive (sub)skills and diagnostic feedback loop that feeds into performance on IELTS speaking tasks and, secondly, explore test-takers' perceptions regarding the use of CDA frameworks.

Literature review

Research on the IELTS speaking test

The primary purpose of the IELTS speaking test is to provide a precise and reliable assessment of the oral proficiency of individuals who are not native English speakers, with the aim of achieving academic and professional objectives (Fernandez, 2018; Taylor, 2011). The assessment consists of an 11–14 min in-person verbal interview with an examiner. The test comprises three components: an introductory conversation, an individual long turn by the test-taker, and a two-way discussion. The current version of IELTS assesses performance based on four key criteria: pronunciation, lexical resource, grammatical range and accuracy, and fluency and coherence. Standardized descriptions are used to describe each band level, which ranges from 1 (i.e., non-user) to 9 (i.e., expert user). The objective of benchmarks and formal scoring rubrics is to optimize uniformity among examiners and testing environments (IELTS, 2024a, b). According to O'Loughlin (2001) and Weir and O'Sullivan (2011), previous studies have demonstrated that inter-rater reliability coefficients for IELTS speaking scores fall within the range of 0.7 to 0.8, generally regarded as excellent. However, research indicates that there is room for enhancement (Peltekov, 2021).

There is a wide range of studies conducted on the IELTS speaking test, examining it from multiple aspects and utilizing various research methodologies. Several research studies (e.g., Ginting et al., 2023; Karim & Haq, 2014; Quaid, 2018; Solihin et al., 2023; Souzandehfar, 2024) have reviewed the IELTS speaking test and provided detailed examinations from various perspectives. For instance, Quaid (2018) and Souzandehfar (2024) examined the authenticity of the IELTS speaking tasks and how they affect the

interactivity of the test. They found that these tasks do not effectively engage or measure the cognitive processes needed for everyday conversation. Souzandehfar (2024) suggests that incorporating problem-solving skills, higher-order thinking skills, and integrated assessment can lead to the development of more authentic and valid L2 tests.

In relation to the subjectivity of raters, Karim and Haq (2014) and Ginting et al. (2023) endeavored to eliminate subjective factors and provided logical and practical recommendations for enhancing the IELTS speaking test. In a separate study conducted in 2023, Solihin et al. sought to provide criticisms of the speaking section of the IELTS test. They questioned the suitability of using interviews as the only method for assessing speaking ability and proposed several measures to address the issues of unfairness, lack of authenticity, and invalidity in the IELTS speaking test results. Regarding the IELTS test, including the speaking section, Read (2022) discussed several concerns, which included the increased emphasis on integrated tasks, the significance of evaluating interactional competence, and the benefits of diagnostic feedback.

Other studies (e.g., Iwashita & Vasquez, 2019; Read & Nation, 2006; Roothooft & Breeze, 2019; Seedhouse et al., 2014) have explored the IELTS speaking rating rubrics. For example, Roothooft and Breeze (2019) aimed to explore grammatical range and accuracy by examining the utilization of both simple and more intricate structures by IELTS candidates across various band levels. They analyzed applicants' attempts at constructing more sophisticated structures, including conditionals, relative clauses, indirect questions, and passive voice across different levels. The analysis revealed a tendency for the utilization of these structures at higher band scores. Similarly, Seedhouse et al. (2014) showed a direct correlation between accuracy, grammatical range, and band score, indicating that a set of assessable speaking characteristics contributes to a specific score on each given IELTS speaking test.

Iwashita and Vasquez (2019) explored the discourse competency qualities seen in test-takers' performances on part 2 of the IELTS speaking test at levels 5, 6, and 7. The aspects of discourse that were evaluated included coherence and cohesive devices. Their findings indicated that coherence and cohesive devices are infrequently observed in the performance of test-takers. Additionally, they observed that discourse devices are utilized in varying ways, but the differences are not statistically significant. These findings suggested that test-takers can make the brief speech necessary for IELTS speaking part 2 more understandable without relying on cohesion and cohesive devices. Nevertheless, disregarding these characteristics could jeopardize the accuracy of the assessment.

With regard to the lexical resource, Read and Nation (2006) analyzed the difficulties of assessing vocabulary as an independent factor in the IELTS speaking test. The researchers discovered that although there were noticeable differences in vocabulary usage between different IELTS band score levels, there was significant variation among test-takers within the same level. This implies that examiners might encounter challenges in reliably evaluating vocabulary performance separately from other subskills when using the current lexical resource criterion. The researchers suggested conducting additional studies to thoroughly enhance the rating scales, providing examiners with better instructions on evaluating vocabulary as an independent factor consistently across different levels of skill in IELTS. According to Souzandehfar (2024), using more advanced techniques of assessment can lead to a test that is more accurate and reliable.

Nakatsuhara (2011) and Schmidgall (2017) have proposed the implementation of a more objective and detailed assessment approach as a means to enhance the validity and reliability of the measurement. An aspect that might be enhanced in the IELTS speaking test is the provision of more comprehensive feedback to test-takers. Currently, candidates are given band scores for each of the four criteria, but they receive limited diagnostic information regarding their particular strengths and weaknesses (Galaczi & Taylor, 2018). Integrating CDA could improve the provision of specific feedback regarding speaking abilities and challenges. The primary objective of CDA is to provide an in-depth understanding of particular skills and knowledge rather than placing exclusive emphasis on a numerical rating (Jang, 2005).

According to Lee and Sawaki (2009), the use of diagnostic models in language assessments has the potential to enhance understanding and feedback by identifying and evaluating specific skill weaknesses. Utilizing CDA in the IELTS speaking test has the potential to allow examiners to precisely identify a candidate's specific areas of deficiency. The inclusion of detailed feedback can assist those taking tests in identifying specific areas for improvement, hence enhancing the effectiveness and efficiency of their preparation (Huhta, 2014). Furthermore, detailed diagnostic information has the potential to improve test-takers' perceptions of the IELTS speaking test. Candidates are likely to see the assessment as being more transparent and equitable if they are provided with detailed and practical feedback rather than solely numerical scores (Xie, 2019). This has the potential to enhance test-taker engagement and drive to improve, given that they can see a direct correlation between the feedback provided and their language learning objectives.

Research on CDA

CDA draws significant inspiration from information processing theory and other theories in the field of cognitive psychology. This theoretical framework posits that the process of learning involves the acquisition of specific cognitive structures and strategies for information processing within a specific domain. According to Sternberg (2011), the human mind is perceived as an information-processing system. CDA is aimed at employing a comprehensive analysis of response patterns in order to elucidate the specific cognitive processes and structures of the individuals being assessed. According to Williamson (2023), the primary objective of CDA is to categorize test-takers into several classes based on their level of proficiency in various qualities or skills being evaluated in the test, rather than providing a single score or grade.

A fundamental assumption posits that a comprehensive understanding of a complex domain necessitates the possession of smaller, discrete skill sets and bodies of knowledge, or attributes (Rupp & Templin, 2008). The procedure involves the identification of certain attributes that an individual possesses or lacks. In recent years, there has been an increasing level of interest in CDA owing to its capacity to offer accurate feedback on various aspects of language knowledge and skills. The purpose of CDA is to ascertain the deficiencies of examinees in certain domain-specific characteristics or their level of competency beyond a general proficiency score.

Drawing on the literature, there are various CDA studies (e.g., Aryadoust, 2012; Effatpanah et al., 2019; Kim, 2010; Mirzaei et al., 2020; Panahi & Mohebbi, 2022; Sawaki et al.,

2009; Xie, 2017) that have made use of CDA for assessing the test-takers' language ability, and they have mainly proved and discussed the efficacy and the diagnostic power of CDA. Early studies in the realm of CDA mostly dealt with receptive L2 skills, especially L2 reading. As a result, there is a growing body of CDA research that has made use of reading tests. For instance, Sawaki et al. (2009) developed a comprehensive diagnostic assessment for reading that establishes a connection between test questions and essential knowledge and skill attributes through the use of a Q-Matrix. The findings provided evidence supporting the Q-Matrix's efficacy in accurately assessing proficiency in several cognitive abilities, including word recognition, identifying main ideas, inferencing, and literary analysis.

In another study in 2020, Mirzaei et al. incorporated the G-DINA (de la Torre, 2011) model into the reading section of IELTS. The researchers generated the final Q-matrix using the reading data of 1025 Iranian pre-IELTS test-takers and the R package CDM, following several rounds that included think-aloud protocol analysis, expert evaluation, and the refinement and validation of the initial Q-matrix. It was postulated that six attributes played a role in the process of IELTS reading. The examinees displayed varying levels of proficiency or deficiency in each of the attributes. CDA has been found to provide an opportunity for IELTS reading test-takers to enhance their areas of weakness and guide L2 teachers and stakeholders in adapting the material that they teach.

With regard to the listening skill, Aryadoust (2012) utilized a cognitive diagnostic model, namely, the fusion model, to analyze the IELTS listening section. By applying the fusion model to the IELTS listening data, the researcher was able to create mastery profiles for test-takers. The paper also addressed the need for more realistic methods of assessing L2 listening and the potential value of CDA in fully representing the construct of L2 proficiency. Likewise, Panahi and Mohebbi (2022) explored the application of CDA to analyze the subskills involved in the IELTS listening test. The authors claimed that the diagnostic information provided by CDA can be beneficial for educational systems to comprehend the underlying structure of the IELTS listening test and to identify L2 learners' strengths and weaknesses in mastering specific subskills. This information might influence language classroom assessment and instruction.

There is a scarcity of research on CDA and writing. In 2010, Kim developed a diagnostic assessment framework through a systematic procedure, resulting in the creation of a descriptor-based diagnostic checklist. This checklist consisted of 35 descriptors that encompassed five distinct writing talents. The researcher examined the accuracy and reliability of the diagnostic data generated by the reduced reparameterized unified model in evaluating the proficiency of individuals in L2 writing. A total of 480 TOEFL essays were assessed by ten L2 educators using the developed checklist. The data analysis provided reliable and accurate differentiation between individuals who possessed advanced skills and those who did not.

Similarly, Effatpanah et al. (2019) and Xie (2017) used the EDD checklist to explore L2 students' writing ability. It was found that CDA could provide detailed diagnostic feedback about the learning status of test-takers. The findings held significant implications for L2 teachers who lacked substantial assessment practices when evaluating the L2 learners' writings. Researchers have also examined CDA in vocabulary assessments. Chen and de la Torre (2013) tested numerous cognitive diagnosis models to classify

word knowledge proficiency in collocation, synonymy, and polysemy. This study discovered that classification models are promising for diagnostic vocabulary assessment.

The present study

There has been a scarcity of CDA studies on speaking, which makes speaking the skill that has received the least amount of investigation in CDA. The gap between the feedback literacy provided in IELTS speaking and the diagnostic methodology employed in assessment necessitates an examination of L2 speaking within the framework of CDA. This study had two primary objectives: (a) to investigate the cognitive subskills required for successful completion of the IELTS speaking test, and (b) to examine the perceptions of IELTS test-takers regarding the IELTS speaking test and the incorporation of CDA into the IELTS assessment. To this end, the following research questions were formulated:

1. What are the underlying cognitive subskills that contribute to successful performance in the IELTS speaking test?
2. What are the perceptions of IELTS test-takers regarding IELTS speaking and the integration of CDA into the IELTS speaking test?

Methodology

Participants

The IELTS speaking test participants included 500 pre-IELTS candidates. Each of them had the intention of attaining a mandatory IELTS overall band score before joining their respective specialty programs. The candidates represented a wide range of backgrounds in terms of age, gender, academic degree, and proficiency level. Among all the test-takers, 151 Iranian pre-IELTS candidates participated in individual interviews to explore their perceptions of the IELTS speaking test (see Table 1 for the test-takers' demographic

Table 1 Test-takers' demographic information

Baseline characteristics	All test-takers	Interviewees
Age		
18–26	173	80
27–35	185	45
36–44	130	24
Over 45	12	2
Gender		
Male	190	91
Female	310	60
Academic degree		
BA	225	89
MA	163	52
PhD	112	18
Proficiency level		
Intermediate	97	11
Upper- intermediate	160	45
Advanced	243	95
Total	500	151

information). Similarly, a group of 20 experts (12 males and 8 females) with 9–30 years of experience in English instruction, specifically in teaching IELTS to Iranian L2 learners, were selected from language institutes located in Isfahan, Iran.

Instruments

IELTS speaking samples

A total of 500 samples of IELTS speaking were gathered. The data set for this study was obtained from many sources, including the IELTS by IDP YouTube channel, the Afarinesh IELTS House YouTube channel, online IELTS mock tests, and the researchers' simulation of the IELTS speaking test. This process was undertaken to ensure the secrecy of the actual IELTS tests.

Think-aloud verbal protocols

Think-aloud verbal protocols, first introduced by Ericsson and Simon (1980, 1993), are widely used in cognitive psychology to understand many aspects of human thinking during a given task. These protocols were meant to collect data and comprehend experts' cognitive processes when diagnosing IELTS-speaking samples. A retrospective approach called verbal protocol analysis was used because the data required note-taking. This method involved experts taking notes and assessing a spoken sample before presenting their opinions. The experts were also asked to rate speaking samples using IELTS band score descriptions.

Interview protocols

Two interviews with a semi-structured format were conducted. The interview conducted to explore the experts' opinions on IELTS speaking samples and verbal protocols, consisted of this set of guiding questions:

1. Did you have any problems with thinking aloud your thoughts and providing comments?
2. What skills or strategies do you think are important in IELTS speaking?
3. How do you provide your students with feedback on their L2 speaking?

During a separate stage of the research, individuals who were preparing for the IELTS test participated in semi-structured interviews that included the following questions:

1. How was your experience with the IELTS speaking test?
2. What did you find the most challenging about the IELTS speaking test?
3. How well do you feel the IELTS speaking band score reflects your true English speaking proficiency?
4. What type of feedback would assist you to gain a better understanding of your speaking abilities and identify areas for improvement?
5. How does knowing your specific ability profile across different IELTS speaking subskills help you plan and direct your future practice and learning?

Procedures

An exploratory qualitative design was utilized to meet the objectives of this study (Creswell, 2018). Think-aloud verbal protocols, interviews, and theme analysis were found to be appropriate for eliciting experts' and test-takers' perceptions as well as identifying potential subskills without being restricted by a priori hypotheses (Braun & Clarke, 2006). A think-aloud verbal protocol was used to record the opinions and comments of 20 IELTS experts on 500 speaking samples. To include all data, the researchers used thematic analysis to record, transcribe, and code verbal processes and follow-up interviews (Braun & Clarke, 2006). The transcriptions were evaluated and sorted into important codes using MAXQDA software (Version 2022) to improve coding reliability and validity. The data was read multiple times to generate codes for IELTS-speaking subskills. Member checking and peer debriefing were used to validate the data (Creswell, 2007; Miles et al., 2014). The experts were asked to assess the selected codes to verify the interpretations. Moreover, external experts conducted peer debriefing to evaluate data analysis and interpretation accuracy. An expert coded 30% of interview transcriptions to test coder agreement and coding reliability. A correlation was established between the experts' IELTS speaking band scores and the original scores to validate their think-aloud techniques for identifying IELTS speaking subskills. After that, Pearson product-moment correlation coefficients were calculated and reported. The experts classified the codes into IELTS-speaking subskills. One of the researchers conducted, recorded, and transcribed 151 semi-structured interviews to answer the second research question. Everyone eagerly joined the research. The participants were assured that their personal data would be kept confidential. Expert-determined band scores and a list of IELTS speaking subskills were offered to the candidates before the interview. Interviewees consented to the audio recording and transcription of oral semi-structured interviews. Version 2022 of MAXQDA software aided codification. Data reliability was supported by member checking and peer debriefing. The 151 test-takers who were interviewed were consulted for member checking. These participants assessed the interpretations' accuracy and reliability by analyzing coded transcripts and emergent themes. An external expert reviewed the data analysis and interpretation techniques during peer debriefing to verify correctness. Additionally, an expert was invited to code 20% of the interview transcriptions.

Results

IELTS speaking subskills

The coding analysis of verbal protocols resulted in the identification of 46 codes as shown in Fig. 1. The total number of code references was 1743, of which *range of vocabulary* (107, 6.02%), *grammar variety* (101, 5.79%), *pronunciation intelligibility* (84, 4.82%), and *complex structures* (82, 4.70%) were the most frequently mentioned IELTS speaking assessment criteria, whereas *past perfect* (11, 0.63%), *infinitives* (11, 0.63%), *singular S* (9, 0.52%), and *gerunds* (8, 0.46%) were the least. Prior to finalizing the codes, 30% of the data was checked by an expert. The calculated inter-coder

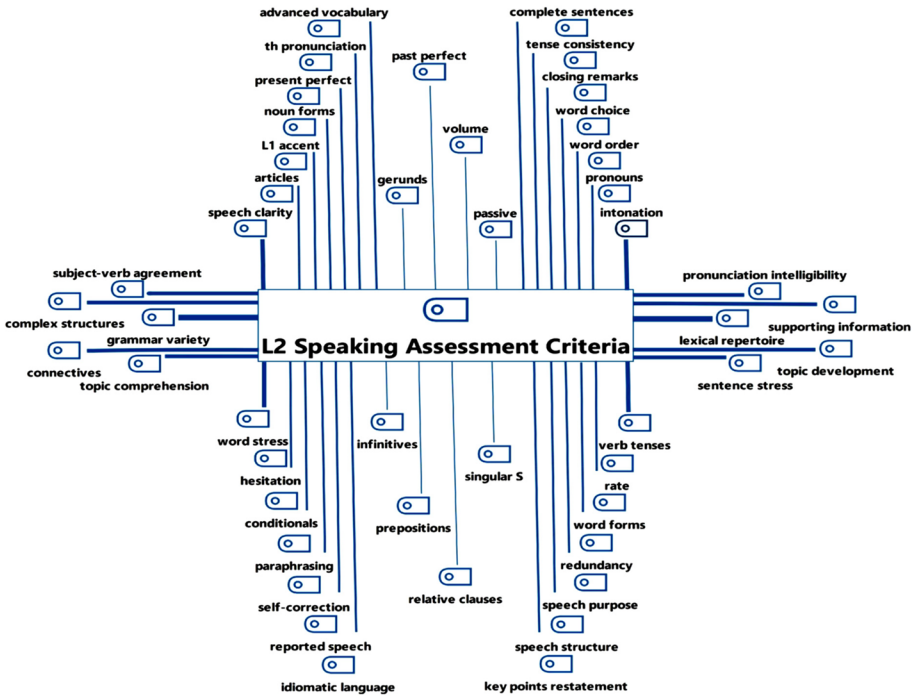

Fig. 1 Factors involved in IELTS *speaking assessment*. *Note*. The thicker the line, the more frequent the criterion

agreement index through Cohen's kappa coefficient turned out to be 92.37%, indicating an acceptable amount of agreement (Miles & Huberman, 1994).

To ensure the reliability of the experts' think-aloud protocols, the IELTS speaking band scores they awarded were correlated with the original IELTS speaking band scores. Pearson product-moment correlation coefficients between pairs of scores ranged from $r = 0.78$ to $r = 0.95$, $p < 0.05$, confirming that the experts' think-aloud verbal protocols were reliable sources for identifying IELTS speaking subskills. The codes were then reviewed by experts in terms of clarity, usefulness, and relevance to IELTS speaking. The experts then grouped the codes into IELTS-speaking subskills in different ways. The examination of the experts' suggested categorizations and the review of the related literature on L2 speaking assessment as well as IELTS speaking assessment by one of the researchers resulted in the identification of five IELTS speaking subskills. Table 2 presents these skills, which were also accepted by all the experts, as well as the definitions of the skills.

IELTS test-takers' perceptions

During the interviews, the participants provided an account of their experience with the IELTS speaking test, elucidated the difficulties encountered, and offered insights into their band score. The written data extracted from interview transcripts were systematically organized, and the relevant data pertaining to the same issue were consolidated.

As shown in Fig. 2, five distinct categories were identified: lack of personalized feedback (133 references), need for multidimensional insights (101 references), bias and subjectivity risks (89 references), limited diagnostic indicators (81), and restrictive

Table 2 IELTS speaking subskills

IELTS speaking subskills	Definition
Linguistic knowledge	It refers to the ability to use a wide range of vocabulary. A speaker who shows strength in this skill uses idiomatic expressions and advanced vocabulary properly. The choice of words is satisfactory, and the speech is clear in terms of meaning.
Syntactic knowledge	It refers to the ability to use a mix of basic and complex grammatical structures. A speaker who shows strength in this skill uses verb tenses and complex structures with no or few mistakes.
Pronunciation mastery	It refers to the intelligible pronunciation of words. A speaker who shows strength in this skill pays special attention to the appropriate use of intonation and also stress in words and sentences. First language has no or little impact on intelligibility. The speech is neither too slow nor too fast, neither too loud nor too soft.
Discourse management and fluency	It refers to the ability to produce extended speech with only occasional hesitation. A speaker who shows strength in this skill answers a specific question or understands and directs a topic in a specific direction using a range of linking words and cohesive devices. The speech is smooth, and the relevant supporting ideas are enough.
Speech structure awareness	It refers to the ability to organize the speech into an introduction, a body, and a conclusion. A speaker who shows strength in this skill starts speaking by stating the purpose of the speech, goes on with a well-structured and informative body, and concludes by summarizing the key points and bringing the speech to an end.

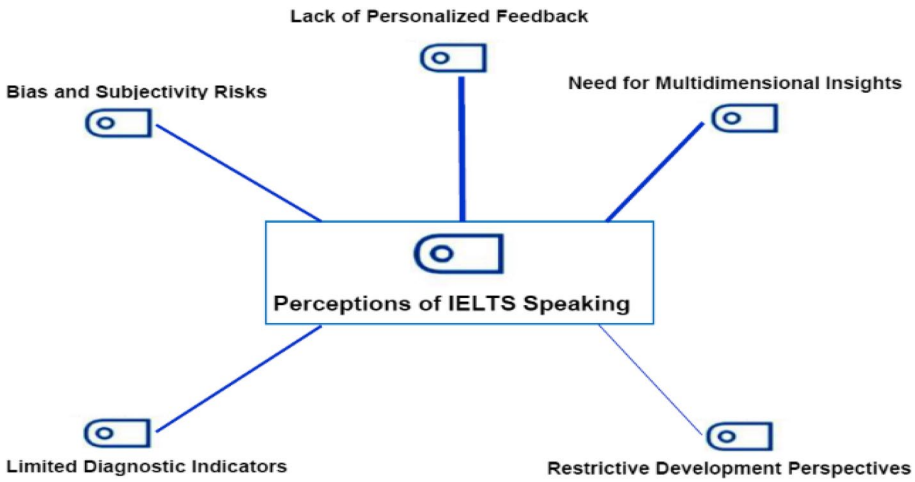

Fig. 2 Test-takers' perceptions of IELTS speaking. *Note*. The thicker the line, the more frequent the criterion

development perspectives (54 references). Table 3 represents these categories and provides examples for each category.

Discussion

In reference to the first research question and the subskills that are associated with the IELTS speaking test, the findings indicated that linguistic knowledge was a significant component in achieving success on the IELTS speaking test. This was in reference to the subskills associated with the IELTS speaking test. According to Storch and Aldosari (2010), linguistic knowledge is defined as the utilization of well-structured, accurate combinations of words, collocations, and pre-established patterns in appropriate

Table 3 Summary of the codification results

Category	Example
Lack of personalized feedback	(1) Personalized techniques based on a thorough examination would allow me to actively direct my own speaking skill improvement. (2) The broad band score and remarks do not assist me to make informed decisions about which specific areas to focus on. (3) The lack of individualized feedback makes me feel as if I don't have control over enhancing my speaking skills in a targeted manner.
Need for multidimensional insights	(1) I wish I could receive more thorough comments on my strengths and weaknesses in the various aspects of speaking proficiency, rather than simply one overall score. (2) The IELTS speaking assessment does not provide adequate insight into my performance across the several characteristics of speaking ability, such as vocabulary, grammar, fluency, and pronunciation. (3) To effectively assess my oral abilities, the IELTS must examine how I blend linguistic, phonological, and fluency factors.
Bias and subjectivity risks	(1) Even with rater training based on comparing assessments, there are intrinsic differences in viewpoints and stringency among raters that might lead to subjectivity. (2) I worry some aspects of my speaking like coherence, grammar or fluency could be over or underweighted by a rater based on their personal biases. (3) Evaluating open-ended speaking performances gives a lot of possibility for rater inconsistency if benchmark alignment and moderation are not strictly enforced.
Limited diagnostic indicators	(1) The IELTS speaking assessment just provides an overall proficiency score, with no diagnostic information about my unique strengths and weaknesses across many dimensions of speaking ability. (2) Assessments of key diagnostic indicators such as vocabulary usage or grammar weaknesses would provide me with extremely valuable self-awareness about the most important areas to focus on. (3) I wish I could obtain more granular feedback that pinpoints the specific skills that require more attention, rather than just a broad classification.
Restrictive development perspectives	(1) My development as a speaker does not follow a straight linear pattern. These comprehensive categories obscure recursive developments as well as variable combinations of strengths and weaknesses. (2) When you rely too much on these restrictive banding structures, you lose the inherent diversity of integrated skill combinations that support overall speaking proficiency. (3) When assessments are restricted to proficiency bands, considerable detail about each student's individual path is lost.

circumstances. This is in contrast to the utilization of fragmented or inaccurate language structures. The research studies by Janebi Enayat and Derakhshan (2021) and Kiliç (2019) support this finding, demonstrating that linguistic knowledge significantly influences L2 performance. However, linguistic knowledge is just one determiner among many in assessing candidates' IELTS speaking proficiency (Read & Nation, 2006).

A further subskill that was identified was syntactic knowledge, which measures the ability of the test-taker to make efficient use of language structures in their responses and conversations. The IELTS band criteria evaluate tense consistency, sophisticated constructions, cohesive devices, clause boundaries, and subordinations that do not contain fragmented simple strings. This finding is similar to Gan's (2012) structural equation modeling study which highlighted grammar subsystems, such as tense forms and grammatical connectedness as key elements in establishing speaking ability. Seedhouse et al. (2014) also revealed that the IETLS candidates' grammatical errors in speaking were reduced as their proficiency level increased.

Mastery of pronunciation was another subskill that was found in the IELTS speaking section. An individual's phonemic articulation, stress patterns, intonation, and interrelated speech features as they are articulated orally are all examined in terms of comprehensibility, clarity, and smoothness when measuring prosodic competency (Isaacs & Trofimovich, 2011). This evaluation is done in order to determine whether or not the individual is proficient in pronouncing. IELTS test-takers with higher levels of skill have a smooth and natural pace, and they do not have any issues with stressing the wrong words or speaking in a manner that requires frequent clarification. Likewise, Dashti and Razmjoo (2020) reported a decrease in the total number of pronunciation errors as the candidates' IELTS band scores improved, suggesting a direct correlation between pronunciation ability and overall level of L2 proficiency.

It was also discovered that fluency and the ability to moderate one's discourse were crucial factors in achieving success on the IELTS speaking test. The test-takers' coherence, speech marking, and interaction abilities are evaluated through the use of discourse management in free-built responses and dialogue simulations (Nakatsuhara, 2011). This type of assessment requires participants to handle subjects, ideas, and turn exchange. Higher bands are distinguished by a coherent and comprehensive main idea, substantial detail in explanation, relevance to the prompt, and a well-defined structure that is reinforced by transitional markers between ideas, as stated by the findings of Galaczi et al. (2011), who analyzed IELTS speaking data using a high-level discourse framework. These features are characterized by higher bands. Clear and uninterrupted speech is a hallmark of proficient speakers. Test-takers in the low band, on the other hand, are more likely to encounter breakdowns while thinking.

Furthermore, it was discovered that the IELTS speaking test assesses the test-takers' awareness of speech structure. This particular subskill seems to be a novel contribution, as it is not explicitly addressed in the traditional subskills of the IELTS speaking test. Speech structure awareness is the cognitive capacity to arrange and structure spoken words in a way that is coherent and understandable when producing spontaneous speech. Oral fluency involves organizing speech into an introduction, a body, and a conclusion. A skilled speaker starts with the speech's aim, then organizes and informs, and ends by briefly highlighting key points and offering closure. Introductions should outline the topic and frame the discourse. The body explains, illustrates, and proves the primary points logically. Conclusions must convey key points and leave a lasting impact (Elsayed, 2017). Beyond grammatical and linguistic proficiency, effective communication is demonstrated by the use of distinct section transitions (Tambunan et al., 2020).

Speech structure awareness might be a more distinct subskill than current models of communicative competence fully account for. The communicative competence model proposed by Canale and Swain (1980) incorporates discourse competence, which encompasses the mastery of cohesion and coherence. However, they do not explicitly mention the tripartite speech structure awareness. Newer models, such as Bachman and Palmer's (1996) organizational knowledge, incorporate textual knowledge, which encompasses cohesiveness and rhetorical organization. However, the explicit subskill of organizing the speech into an introduction, body, and conclusion is not emphasized. Our discovery of speech structure awareness indicates that this particular organizational subskill might be more cognitively distinct and rhetorically

significant than current models acknowledge. The identification of this subskill necessitates a re-evaluation of how we assess oral proficiency, particularly in high-stakes assessments like IELTS. By integrating this particular subskill into the scoring rubrics for IELTS speaking, evaluators can enhance the consistency and reliability of their assessments regarding a speaker's level of proficiency. This has the potential to decrease the subjective and variable nature of scoring, resulting in enhanced reliability and fairness in the evaluation process.

With regard to the second research question and the perceptions of the IELTS speaking test, IELTS test-takers frequently voiced their dissatisfaction with the IELTS's existing ways of evaluating speaking skills, stating that they lack the ability to provide personalized feedback that is tailored to the test-takers' individual strengths and weaknesses. Garin (2019) also demonstrated excellent proof that detailed feedback has the potential to improve IELTS speaking accuracy. According to Huhta et al. (2006), test-takers are provided with a broad proficiency band score and qualitative feedback rather than development methods that are specifically targeted to their needs. Students are unable to establish measurable micro-progress targets that are appropriate for their development if they do not have access to individualized benchmarking data that provides a thorough explanation of how to achieve specific levels of ability. Lack of clarity prohibits examinees from making informed choices about their personal interests.

The findings also identified the need for multidimensional insights and revealed that many test-takers thought that the IELTS speaking test did not really measure how well someone could speak, considering subskills like linguistic knowledge, syntactic knowledge, pronunciation mastery, discourse management and fluency, and speech structure awareness. Based on Read (2022), IELTS has demonstrated a strong resistance to any significant reconsideration of L2 competence, including more emphasis on integrated tasks and the requirement to assess interactional competence. Speaking well requires many subskills, and tests that solely rely on total band scores and overall rubrics lack sufficient detail (Papageorgi et al., 2010). There are linguistic, phonological, and fluency factors in L2 speaking. People taking tests have a hard time judging their own success because there are not any clear-cut examples or calibrated continuum tests that show the different levels of proficiency for each speaking ability. The best way to test these traits is with multiple diagnostic methods that look at both general and specific skills.

Several test-takers noted concerns about IELTS scoring reliability and consistency, which might be due to subjective biases in existing qualitative methodologies used to assess speaking proficiency. According to Ginting et al. (2023), the IELTS speaking test is delivered by a single rater, which means that the interviewer's subjectivity and unpredictability influence a candidate's declared skill level. Because just one rater administers the IELTS speaking test, this level could not match the candidate's innate ability. Specifically, rater viewpoints, experience, stringency, and other characteristics create subjectivity concerns in assessments (Suen, 2014). There are higher validity risks when you look at integrated performance-based tasks like grading spontaneous speech which shows how hard language is in real life. Raters might use their own biases to decide how much to value quality traits like coherence, syntax, and fluency, instead of following established rubrics. To eliminate subjectivity, Karim and Haq (2014) suggested assessing the IELTS speaking test by two separate raters who are not aware of each other's scoring methods.

Another criticism was the lack of valuable diagnostic indicators that could reveal the strengths and weaknesses in the speaking proficiency of IELTS test-takers. It is possible that values that only prioritize broad competency categories will ignore special abilities and micro skills that require individualized treatment (Papageorgi et al., 2010). This finding is congruent with Solihin et al.'s (2023) findings, which criticized using interviews as the only method to assess speaking proficiency. IELTS seems to not fully cover the wide range of language skills used in real-life communication situations (Nunan, 1999).

According to Solihin et al. (2023), IELTS seems to primarily measure the formal register of language usage through interviews. However, this fails to adequately depict real-life settings, whereby both formal and informal linguistic types are applied, and the language employed across contexts changes greatly. This gap becomes apparent when analyzing the two different modes of the IELTS exam, specifically the general training and academic versions. Significantly, the speaking section does not distinguish between the two separate settings. Therefore, IELTS might not accurately assess or evaluate speaking proficiency in its most authentic form. In addition, although interviews are a common method for evaluating speaking talents, over-reliance on this method might lead to the neglect of other crucial speaking skills that are necessary in various situations.

Last but not least, placing too much emphasis on standardized proficiency levels as a way to understand the growth, complexity, and depth of IELTS speaking competence can have controlling effects. Focusing on complete banding systems can leave out skill development that goes beyond classification. Authentic assessment evaluates test-takers' proficiency in real-life tasks, showcasing their meaningful utilization of fundamental knowledge and abilities (Quaid, 2018; Souzandehfar, 2024). However, standardized tests inadequately capture the complex relationship among various skills, which could limit students' ability to communicate effectively in real-life situations. (McNamara & Knoch, 2012). Despite the strong correlation between communication ability and academic performance, Curtis (2004) suggests using multidimensional evaluations to assess test-takers' proficiency levels.

Without a doubt, the significant shifts in perception move the focus of language assessment from standardization to recognizing and appreciating individual skill profiles. L2 learners show readiness by emphasizing their abilities rather than rigorous standards. One examination can narrow the value of diversity. CDA handles these challenges statistically by breaking capability into hierarchically modeled components. Offering clear and direct instruction on IELTS speaking subskills, especially the speech structure awareness as presented in this research, can aid learners in cultivating these crucial subskills, ultimately improving their overall speaking proficiency and performance in significant assessments such as IELTS. All the subskills that have been identified can serve as the foundation for constructing a Q-matrix and developing cognitive diagnostic models in future studies.

Implications

The findings of the study offer significant insights into the cognitive processes involved in measuring speaking constructs during the IELTS test. The research elucidates the multi-faceted character of oral proficiency by defining the precise linguistic knowledge, abilities, and tactics that contribute to successful speaking performance. The findings of

this study have significant implications for enhancing the assessment of speaking subskills, not only in the context of IELTS but also in other standardized language competence assessments. Furthermore, the analysis conducted in the study aids in determining the practicality and effectiveness of utilizing CDA to improve the interpretation and feedback on IELTS speaking.

The provision of precise feedback can effectively steer training that is more focused and detailed to meet the unique needs identified by learners. Diagnostic assessment enables the identification of specific deficits in subskills, allowing for the customization of educational priorities based on data rather than using broad approaches. This pedagogical approach effectively utilizes evaluation as a means to augment the process of learning. CDA can significantly enhance the interpretability and pedagogical usefulness of high-stakes speaking tests by examining the cognitive underpinnings of speaking ability and providing feedback at the skill level. In addition to advancing the field of language assessment as a whole, the positive results of this study are a crucial step toward putting these advantages into practice.

The results are significant for Iran, as the IELTS test is a crucial English language competency examination for academic and professional admission. Iranian test-takers can obtain more comprehensive diagnostic information regarding their specific areas of strength and weakness in L2 speaking through CDA. This can assist individuals in honing their language acquisition skills and preparing for the IELTS examination. Given the significant importance of IELTS in Iran, providing cognitive diagnostic information could help reduce stress and uncertainty by providing learners with a clear understanding of their current standing and the specific subskills they need to enhance.

In fact, Iranian students can enhance their autonomy in learning through the use of CDA reporting for self-directed study and test preparation. Moreover, IELTS preparation programs and language centers in Iran might utilize CDA models at an institutional level to develop detailed and data-driven curricula and interventions that specifically target the speaking subskills that students need to enhance. Instead of a standardized training approach, CDA would enable more personalized learning paths. This aligns with the increasing focus on diagnostic and personalized instruction in language education.

Conclusion

The lack of diagnostic information in the speaking score poses challenges and unfairness for test-takers when making high-stakes decisions. Enhancing learning and increasing the educational value of the evaluation could be achieved by offering more tailored feedback. Modifying IELTS speaking test ratings is necessary due to their limited diagnostic use. The primary objective of this study was to ascertain the cognitive subskills that contribute to the successful completion of the IELTS speaking test. Subsequently, the study aimed to investigate the perceptions of IELTS test-takers towards the IELTS speaking test and the incorporation of CDA into the assessment of IELTS. This study revealed five distinct subskills, namely, linguistic knowledge, syntactic knowledge, pronunciation mastery, discourse management and fluency, and speech structure awareness. CDA could be used as an aid tool to evaluate linguistic knowledge, syntactic knowledge, pronunciation mastery, discourse management and fluency, and speech structure awareness

to assess speaking proficiency, give test-takers more detailed feedback, and improve the IELTS speaking test.

CDA's secondary diagnostic information can improve the IELTS band score, allowing for a more thorough assessment of speaking proficiency. Test-takers who just missed a goal band score might find areas that require further practice with the additional information. CDA helps test providers verify the IELTS speaking test's correctness and offer performance data. This strategy simplifies the examination process and allows test-takers to pursue specialized learning pathways, enhancing their chances of attaining their language proficiency goals. The addition of CDA to the IELTS speaking test could improve and authenticate the evaluation. By examining patterns and trends among a broad pool of test-takers, researchers and test producers can refine the test, change the scoring rubrics, and guarantee it tests the intended components. CDA in the IELTS speaking test can improve assessment, provide detailed feedback, and provide targeted learning opportunities for test-takers.

When applied to the Iranian context, the implementation of CDA has the potential to have a significant influence on L2 teaching, learning, and evaluation practices, particularly in relation to the high-stakes assessment of IELTS for academic and professional purposes. CDA has the potential to reduce the anxiety and ambiguity associated with the IELTS test by providing thorough cognitive diagnostic feedback on the test-takers' speaking performance. CDA would offer detailed insights into the precise language skills and subskills that the learner has learned or needs to improve upon, in contrast to typical assessments that just offer an overall score.

With diagnostic feedback in hand, Iranian students could take a more targeted and personalized approach to their preparation for IELTS, concentrating on the specific areas in which they need to improve. Consequently, learners would avoid expending time and energy on subskills that they have already mastered, resulting in more efficient and effective learning. Overall, the implementation of CDA in the Iranian context has the potential to revolutionize the way in which IELTS speaking is taught, learned, and evaluated. This would result in a more personalized, targeted, and transparent approach to language learning and evaluation.

Current research limitations necessitate more research. First, IELTS speaking subskills were identified using think-aloud verbal data from experts, focusing on important factors. This approach did not accurately represent test-takers' speaking knowledge, procedures, and tactics during their speaking activities. Testing these speaking processes with students' retrospective verbal protocols might have more accurately identified the IELTS speaking subskills and speaking abilities for evaluation. The current body of literature on scale development lacks sufficient research that supports the incorporation of test-takers' perceptions; therefore, additional investigation is required in this domain.

Furthermore, only Iranian IELTS test-takers were interviewed to address the second research question due to accessibility issues. A possible expansion of this work could examine IELTS speaking perceptions in other scenarios to improve comprehension. In addition, this study examines the feasibility and value of CDA for the IELTS speaking assessment. However, more validation research is needed to create a Q-matrix from the current study's data. As a diagnostic speaking checklist, the Q-matrix will establish a correlation between the subskills and performance on the IELTS speaking test. It is

essential to develop such a robust diagnostic instrument in order to establish firmer psychometric justifications for CDA-based scoring methodologies and to provide more precise feedback. In fact, this would enable the triangulation of both qualitative and quantitative evidence, facilitating a more rigorous and complete validation of the technical features and score interpretations of the CDA framework.

Abbreviations

CDA Cognitive Diagnostic Assessment
IELTS International English Language Testing System
L2 Second language

Acknowledgements
The authors would like to extend their sincere appreciation to all the experts who kindly helped with data collection and analysis in this study.

Authors' contributions
MHV conceptualized the study, collected and analyzed the data, and wrote the first draft of the manuscript. AM reviewed and revised the original draft. AR read the manuscript and contributed to its consistency and coherence. All authors read and approved the final manuscript.

Authors' information
MHV is a Ph.D. student of TEFL at Shahrekord University. She has publications in ISI-indexed journals such as Computer Assisted Language Learning, Studies in Educational Evaluation, and Human Arenas. Her area of interest includes Language Testing and Assessment, Educational Psychology, Teacher Education, and Second Language Learning.
AM is an associate professor of Applied Linguistics at Shahrekord University. He has numerous publications in different reputable journals (e.g., System, ReCALL, Journal of Pragmatics, Studies in Educational Evaluation, Educational Psychology, CALL, etc.). His research interests include Computer-Assisted Language Learning, Vygotsky-inspired Sociocultural Theory, Interlanguage and Intercultural Pragmatics, and Language Testing/Assessment.
AR is an associate professor of applied linguistics in the English Department at Shahrekord University in Iran. His area of interest includes educational psychology, language testing, and textbook evaluation. He is very interested in research on affective variables in second/foreign (L2) learning/teaching and computer-assisted language learning (CALL). He has published 5 books and 92 papers, including papers in ISI-indexed journals such as Computer Assisted Language Learning, Language Learning & Technology, Language Teaching, and Journal of Multilingual and Multicultural Development.

Funding
Not applicable.

Availability of data and materials
The datasets generated and analyzed during the current study are not publicly available as noted in the privacy agreement with the participants.

Declarations

Competing interests
The authors declare no competing interests.

Received: 6 May 2024 Accepted: 5 September 2024
Published online: 12 October 2024

References

Allen, D. (2016). Investigating washback to the learner from the IELTS test in the Japanese tertiary context. *Language Testing in Asia, 6*(7), 1–20. https://languagetestingasia.springeropen.com/articles/10.1186/s40468-016-0030-z
Aryadoust, V. (2012). Using cognitive diagnostic assessment to model the underlying structure of a listening test: A subskill-based approach. *The Asian EFL Journal, 14*(4), 81–106.
Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford University Press.
Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa
Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47. https://doi.org/10.1093/applin/I.1.1
Celce-Murcia, M. (2007). Rethinking the role of communicative competence in language teaching. In E. A. Soler & M. P. S. Jordà (Eds.), *Intercultural language use and language learning* (pp. 41–57). Springer.
Chen, J., & de la Torre, J. (2013). A general cognitive diagnostic modeling approach for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*(6), 419–437. https://doi.org/10.1177/0146621613479818

Clark, T., Holland, M., & Spiby, R. (2022). Seeking empirical evidence to support online test validation. In K. Sadeghi (Ed.), *Technology in language assessment* (pp. 14–31). Routledge.
Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). Sage.
Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage.
Crystal, D. (2012). *English as a global language* (2nd ed.). Cambridge University Press.
Curtis, D. D. (2004). The assessment of generic skills. In J. Gibb (Ed.), *Generic skills in vocational education and training: Research readings* (pp. 136–156). National Centre for Vocational Education Research.
Dashti, L., & Razmjoo, S. A. (2020). An examination of IELTS candidates' performances at different band scores of the speaking test: A quantitative and qualitative analysis. *Cogent Education, 7*(1), 1–27. https://doi.org/10.1080/2331186X.2020.1770936
de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7
Derakhshan, A., & Kordjazi, M. (2015). Implications of dynamic assessment in second/foreign language contexts. *English Linguistics Research, 4*(1), 41–48. https://doi.org/10.5430/elr.v4n1p41
Deygers, B., Van Gorp, K., & Demeester, T. (2018). The b2 level and the dream of a common standard. *Language Assessment Quarterly, 15*(1), 44–58.
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
Elsayed, A. M. M. (2017). *Developing EFL student-teachers' oral communication skills in light of the Toastmasters approach*. Master's thesis, Ain Shams University.
Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. MIT Press.
Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215–251. https://psycnet.apa.org/doi/10.1037/0033-295X.87.3.215
Fernandez, C. J. (2018). Behind a spoken performance: Test takers' strategic reactions in a simulated part 3 of the IELTS speaking test. *Language Testing in Asia, 8*(1), 1–20. https://languagetestingasia.springeropen.com/articles/10.1186/s40468-018-0073-4
Galaczi, E. D., Ffrench, A., Hubbard, C., & Green, A. (2011). Developing assessment scales for large-scale speaking tests: A multiple-method approach. *Assessment in Education: Principles, Policy & Practice, 18*(3), 217–237. https://doi.org/10.1080/0969594X.2011.574605
Galaczi, E. D., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly, 15*(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816
Gan, Z. (2012). Understanding L2 speaking problems: Implications for ESL curriculum development in a teacher training institution in Hong Kong. *Australian Journal of Teacher Education, 37*(1), 43–59. https://doi.org/10.14221/ajte.2012v37n1.4
Garin, A. (2019). Improving speaking fluency for the IELTS test. In B. Dubin, M. T. Nguyen, & T. Past (Eds.), *Classroom research in the Japanese context* (pp. 1–9). Temple University.
Ginting, R. S., Dalimunte, A. A., Dalimunte, M., Kurniati, E. Y., & Adelita, D. (2023). A critical review of IELTS speaking test. *JL3T (Journal of Linguistics, Literature and Language Teaching), 9*(2), 138–155. https://doi.org/10.32505/jl3t.v9i2.7161
Hamid, M. O., Hardy, I. & Reyes, V. (2019). Test-takers' perspectives on a global test of English: Questions of fairness, justice and validity. *Language Testing in Asia, 9*(16). https://doi.org/10.1186/s40468-019-0092-9
Huhta, A. (2014). Diagnostic and formative assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 775–791). Wiley-Blackwell.
Huhta, A., Kalaja, P., & Pitkänen-Huhta, A. (2006). Discursive construction of a high-stakes test: The many faces of a test-taker. *Language Testing, 23*(3), 326–350. https://doi.org/10.1191/0265532206lt331oa
Hymes, D. H. (1972). On Communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Penguin.
IELTS. (2024a). *IELTS scoring in detail*. Retrieved April 2, 2024, from https://ielts.org/organisations/ielts-for-organisations/ielts-scoring-in-detail
IELTS. (2024b). *IELTS test format explained*. Retrieved April 2, 2024, from https://takeielts.britishcouncil.org/take-ielts/test-format
Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics, 32*(1), 113–140. https://doi.org/10.1017/S0142716410000317
Iwashita, N., & Vasquez, C. (2019). An examination of discourse competence at different proficiency levels in IELTS speaking part 2 proposals. *IELTS Research Reports, 5*, 1–44. https://s3.eu-west-2.amazonaws.com/ielts-web-static/production/Research/examination-of-discourse-competence-at-different-proficiency-levels-in-ielts-speaking-part-2-iwashita-et-al-2015.pdf
Janebi Enayat, M., & Derakhshan, A. (2021). Vocabulary size and depth as predictors of second language speaking ability. *System, 99*(3), 1–15. https://doi.org/10.1016/j.system.2021.102521
Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. Doctoral dissertation, University of Illinois at Urbana-Champaign.
Karim, S., & Haq, N. (2014). An assessment of IELTS speaking test. *International Journal of Evaluation and Research in Education (IJERE), 3*(3), 152–157. https://doi.org/10.11591/ijere.v3i3.6047
Kiliç, M. (2019). Vocabulary knowledge as a predictor of performance in writing and speaking: A case of Turkish EFL learners. *PASAA, 57*, 133–164. https://doi.org/10.58837/CHULA.PASAA.57.1.6
Kim, Y. H. (2010). *An argument-based validity inquiry into the empirically-derived descriptor-based diagnostic (EDD) assessment in ESL academic writing*. Doctoral dissertation, University of Toronto.
Kurata, N. (2011). *Foreign language learning and use: interaction in informal social networks*. Continuum International Publishing Group.

Lee, Y. W., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly, 6*(3), 172–189. https://doi.org/10.1080/15434300902985108
Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.
Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.
McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555–576. https://doi.org/10.1177/0265532211430367
Miles, M. B., & A. M. Huberman. (1994). *Qualitative data analysis: An expanded sourcebook*. Sage.
Miles, M. B., A. M. Huberman, & J. Saldana. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Sage.
Minakova, V. (2020). Dynamic assessment of IELTS speaking: A learning-oriented approach to test preparation. *Language and Sociocultural Theory, 6*(2), 184–212. https://doi.org/10.1558/lst.36658
Mirzaei, A., Heidari, M. V., & Hashemian, M. (2020). Retrofitting the IELTS reading section with a general cognitive diagnostic model in an Iranian EAP context. *Studies in Educational Evaluation, 64*, 1–10. https://doi.org/10.1016/j.stueduc.2019.100817
Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28*(4), 483–508. https://doi.org/10.1177/0265532211398110
Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. D. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly, 14*(1), 1–18. https://doi.org/10.1080/15434303.2016.1263637
Nunan, D. (1999). *Second language teaching & learning*. Heinle & Heinle
O'Loughlin, K. (2001). *Studies in language testing 13: The equivalence of direct and semi-direct speaking tests*. Cambridge University Press.
Panahi, A., & Mohebbi, H. (2022). Cognitive diagnostic assessment of IELTS listening: Providing Feedback from its internal structure. *Language Teaching Research Quarterly, 29*, 147–160. https://doi.org/10.32038/ltrq.2022.29.10
Papageorgi, I., Haddon, E., Creech, A., Morton, F., De Bezenac, C., Himonides, E., Potter, J., Duffy, C., Whyton, T., & Welch, G. (2010). Institutional culture and learning II: Inter-relationships between perceptions of the learning environment and undergraduate musicians' attitudes to performance. *Music Education Research, 12*(4), 427–446. https://doi.org/10.1080/14613801003746550
Peltekov, P. (2021). The international English language testing system (IELTS): A critical review. *JELTL, 6*(2), 395–406. https://doi.org/10.21462/jeltl.v6i2.581
Quaid, E. D. (2018). Reviewing the IELTS speaking test in East Asia: theoretical and practice-based insights. *Language Testing in Asia, 8*(2), 1–9. https://languagetestingasia.springeropen.com/articles/10.1186/s40468-018-0056-5
Read, J., & Nation, P. (2006). An investigation of the lexical dimension of the IELTS speaking test. *IELTS Research Reports, 6*, 1–25. https://s3.eu-west-2.amazonaws.com/ielts-web-static/production/Research/investigation-oflexical-dimension-of-ielts-speaking-test-read-et-al-2006.pdf
Read, J. (2022). Test review: The international English language testing system (IELTS). *Language Testing, 31*(4), 679–694. https://doi.org/10.1177/02655322221086211
Roothooft, H., & Breeze, R. (2019). IELTS: Investigating the development of 'grammatical range and accuracy' at different proficiency levels in the IELTS speaking test. *IELTS Research Reports, 1*, 2–36. https://s3.eu-west-2.amazonaws.com/ielts-web-static/production/Research/investigating-development-of-grammatical-range-and-accuracy-at-different-proficiency-levels-roothooft-et-al-2019.pdf
Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives, 6*(4), 219–262. https://psycnet.apa.org/doi/10.1080/15366360802490866
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*(3), 190–209. https://doi.org/10.1080/15434300902801917
Schmidgall, J. E. (2017). The consistency of TOEIC® speaking scores across ratings and tasks. *ETS Research Report Series, 2017*(1), 1–8. https://doi.org/10.1002/ets2.12178
Seedhouse, P., Harris, A., Naeb, R., & Ustunel, E. (2014). The relationship between speaking features and band descriptors: a mixed methods study. *IELTS Research Reports, 2*, 1–30. https://ielts.org/researchers/our-research/research-reports/the-relationship-between-speaking-features-and-band-descriptors-a-mixed-methods-study
Solihin, S., Utami, D. R., Aprianti, A. D. I., & Mayang Sari, M. (2023). Issues with the speaking section of the international English language testing system (IELTS): An assessment critique. *ELS Journal on Interdisciplinary Studies in Humanities, 6*(3), 540–545. https://doi.org/10.34050/elsjish.v6i3.28355
Souzandehfar, M. (2024). New perspectives on IELTS authenticity: an evaluation of the speaking module. *International Journal of Language Testing, 14*(1), 34–55. https://doi.org/10.22034/ijlt.2023.409599.1272
Sternberg, R. J. (2011). *Cognitive psychology* (6th ed.). Cengage Learning.
Storch, N., & Aldosari, A. (2010). Learners' use of first language (Arabic) in pair work in an EFL class. *Language Teaching Research, 14*(4), 355–375. https://doi.org/10.1177/1362168810375362
Suen, H. K. (2014). Peer assessment for massive open online courses (MOOCs). *The International Review of Research in Open and Distributed Learning*, *15*(3), 312–327. https://www.learntechlib.org/p/148201/
Tambunan, V. D., Raja, V. L., & Sari, A. S. P. (2020). Improving students' speaking skill through story telling technique to the eleventh grade students of SMA Swasta Katolik Budi Murni 2 Medan. *Kairos ELT Journal, 2*(1), 28–46. https://doi.org/10.54367/kairos.v2i1.723
Taylor, L. (2011). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge University Press.
Taylor, L., & Jones, N. (2006). Cambridge ESOL exams and the common European framework of reference (CEFR). *Research Notes, 24*, 2–5.
Weir, C. J., & O'Sullivan, B. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: Theories and practices: Palgrave advances in language and linguistics.* Palgrave Macmillan.

Williamson, J. (2023). *Cognitive diagnostic models and how they can be useful*. Cambridge University Press & Assessment. https://www.cambridgeassessment.org.uk/Images/701443-cognitive-diagnostic-models-and-how-they-can-be-useful.pdf

Wu, X., Zhang, L. J., & Liu, Q. (2021). Using assessment for learning: Multi-case studies of three Chinese university English as a foreign language (EFL) teachers engaging students in learning and assessment. *Frontiers in Psychology, 12*, 1–15. https://doi.org/10.3389/fpsyg.2021.725132

Xie, Q. (2017). 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

Xie, Q. (2019). Diagnosing linguistic problems in English academic writing of university students: An item bank approach. *Language Assessment Quarterly: An International Journal, 17*(1), 1–31. https://doi.org/10.1080/15434303.2019.1691214

Publisher's Note