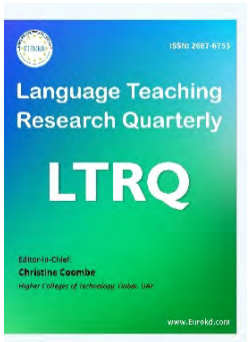


Language Teaching Research Quarterly

2023, Vol. 37, 231–247

Mixed Methods Investigation into Test Score Users' Perspectives about IELTS Reading Skill Profiles

Eunice Eunhee Jang*, Christie Barron, Hyunah Kim, Bruce Russell

Ontario Institute for Studies in Education, University of Toronto, Canada

Received *14 May 2023* ***Accepted*** *21 October 2023*

Abstract

Research on the use of standardized test scores in higher education reveals significant variations in attitudes and perceptions of language proficiency tests among test score users. Most test score users have limited knowledge about test score interpretations in terms of what English as additional language (EAL) students typically know and can do at the language proficiency levels associated with admission cut scores. To address this critical gap, the language testing field has actively investigated the potential of Diagnostic Classification Models (DCMs) to offer useful information, facilitating test score users in their decision-making processes. The present two-phase mixed methods study examined the characteristics of reading skill profiles across various IELTS band scores, specifically focusing on the most frequently used admission cut scores: 6.0, 6.5, and 7.0. The study further explored test score users' perspectives about these admission test scores, challenges encountered by EAL students, and the usefulness of reading skill profiles derived from DCMs. Findings from the application of DCMs to IELTS reading test responses ($N = 5{,}222$) showed a lack of advanced skills, such as inferential reasoning, at these commonly employed cut scores. Test score users perceived the skill profiles as instrumental in distinguishing reading abilities across various band scores and discussed the EAL students' lack of critical reasoning, especially in inferential and synthesis tasks in coursework. The results underscore the potential of DCM-based skill profiles in providing test score users with detailed information about test score interpretations.

Keywords: *Score Interpretations, Admission Cut Scores, EAL Students, Diagnostic Classification Modeling, IELTS*

Prologue

In his insightful book (2014), *Mixed methods research for TESOL*, Brown references my dissertation research (Jang, 2005) as an example of guided reading. In that study (2005), I examined the implications of cognitive diagnostic testing on teaching and learning by focusing on diagnostic assessment's construct definition, psychometric modeling, and feedback

* Eunice Eunhee Jang
E-mail address: eun.jang@utoronto.ca
https://doi.org/10.32038/ltrq.2023.37.13

utilization. Brown remarks, “I would guess that at least a couple of her articles owe something to her dissertation research. However, this scholar has clearly developed well beyond her original dissertation research. I hope that Professor Jang will serve as an inspiration to you all” (Brown, 2014, p. 218). In recent empirical mixed methods research co-authored with my graduate students, we further delve into refining score interpretations through cognitive diagnostic skill profiling for test score users. We hope our work contributes to Brown’s dedication to progress language testing and assessment through rigorous psychometric methodologies and comprehensive mixed methods inquiries.

Introduction

Contemporary views on test validity prioritize the significance of meaningful test score interpretations and the positive effects of test score use (AERA et al., 2014; Bachman & Palmer, 2010; Chapelle et al., 2008; Kane, 2006; Taylor & Weir, 2012). This perspective holds significant implications for current language testing practices, especially in the context of admitting international English-as-an-additional-language (EAL) students to post-secondary programs in English-medium universities. Given the crucial role of English language proficiency testing in these admissions, it is imperative to support test score users with clear test score interpretations and defensible decision-making processes tailored to their local contexts (Milanovic & Weir, 2010). For various purposes, including admissions and language program placements, test score users need deeper insights into what test takers can typically achieve at different proficiency levels (Hyatt & Brooks, 2006; Ingram & Bayliss, 2007). This renewed understanding of test validity has encouraged the fields of language testing and educational measurement to investigate ways to provide useful information for test score users, facilitating their decision-making processes and resource allocations (Lee & Sawaki, 2009).

Within this framework, diagnostic classification modeling (DCM) approaches (DiBello & Stout, 2007; Rupp et al., 2010) have garnered significant attention over the past two decades, with the objective of addressing such demands. DCMs begin with a clear specification of how different items align with distinct sets of skills (or attributes), often specified in a Q-matrix. The primary goal of DCMs is to classify individual test takers into discrete multidimensional skill profiles, based on a pre-defined Q-matrix.

Despite the evident potential of DCMs, most score reporting practices still predominantly use a unidimensional scale. In terms of language prerequisites for post-secondary admissions, multiple cut-off scores are generally set for different subtests based on modality, as well as an overall score. However, they do not offer the fine-grained skill profiling DCMs offer. Test score users have voiced the need for more detailed information that could facilitate meaningful score interpretations and programming (Hyatt & Brooks, 2006; Ingram & Bayliss, 2007).

In the current study, we examined the characteristics of reading comprehension skill profiles across various International English Language Test System (IELTS) band scores, a standardized English proficiency test widely utilized for post-secondary admissions. Our aim was to explore DCMs in order to harness the potential of skill profiling within the context of post-secondary educational settings where high-stakes standardized tests are widely used for admissions. We also engaged test score users to examine how they interpret and utilize reading skill profiles in their admissions decisions and support programming.

Literature Review

The field of language testing and assessment has witnessed a paradigm shift toward effects-driven assessment practices via diagnostic feedback (Lee, 2015). DCM approaches have been extensively researched and applied to language tests measuring reading (e.g., Jang, 2005, 2009; Jang et al., 2015; Kim, 2015; Li et al., 2016), listening (e.g., Dong et al., 2021; Min & He, 2022; Yi, 2017), grammar (Clark & Endres, 2021), and writing (e.g., Effatpanah, 2019; Xie, 2017; Zhai et al., 2022). Such significant attention to the potential of DCM reflects desires and pressures to provide helpful information for test score users, aiding their decision-making processes and resource allocations for student support (von Davier & Lee, 2019).

Existing DCMs differ based on their a priori assumptions about inter-skill relationships. General DCMs, such as the General Diagnostic Model (GDM) by von Davier (2005) and the Generalized Deterministic-Input Noisy-and-Gate Model (G-DINA) by de la Torre (2011), do not require an a priori specification of the inter-skill relationship, whether compensatory or conjunctive. Instead, these models allow both types to fit items within the same test (Ravand, 2015; Rupp & Templin, 2008). In contrast, specific DCMs, such as the Deterministic-Input, Noisy-or-Gate Model (DINO) by Templin and Henson (2006) and the Compensatory Reparameterized Unified Model (C-RUM) by Hartz et al. (2002), assume compensatory relationships. In these models, a deficit in mastery of one skill can be compensated for by proficiency in another for successful item performance. On the other hand, conjunctive DCMs, such as the Reparameterized Unified Model (RUM) by DiBello et al. (1995) and the Deterministic-Input, Noisy-and-Gate Model (DINA) by Junker and Sijtsma (2001), presuppose that a correct item response requires mastery of all specified skills. The quality of diagnostic information derived from a DCM depends on comprehensive specifications of linguistic knowledge and cognitive skills elicited by test items. Its efficacy for guiding test score interpretations and use depend on how well the elicited skills represent real-life language demands that impact students' academic performance.

The increasing demand for comprehensive information from large-scale assessment programs has prompted the practice of applying retrofitting analyses to existing large-scale tests, even those not initially intended for diagnostic purposes (Sessoms & Henson, 2018). To address these limitations, efforts have been made to enhance the diagnostic utility of retrofitted tests (e.g., Kim, 2015; Min & He, 2022). For instance, Kim (2015) applied a reduced reparameterized unified model (Hartz et al., 2002) with ten skills to university placement reading test scores for an adult English-as-a-second-language program. The study's goal was to investigate methods of providing diagnostic feedback to university stakeholders. The study findings demonstrated how reporting skill mastery probabilities, categorized by low, medium, and high proficiency levels, could be used for curriculum refinement and skill-profile-informed instruction.

Alternatively, more recent research has embraced DCM by designing and implementing diagnostic assessments based on multidimensional cognitive theories of language ability (Min et al., 2022; Toprak & Cakir, 2021). For example, Toprak and Cakir (2021) employed a multi-step field study to inductively develop an English-as-a-foreign-language diagnostic assessment with five higher-order reading skills for academic settings in Turkey. Following this

development, they analyzed student responses using a general log-linear cognitive diagnosis model and provided diagnostic reports to the study participants.

Though skill mastery profiles estimated from DCMs offer comprehensive insights into the particular skills students have mastered or not, multidimensional item response theory models are infrequently used for item calibration and score reporting in practice. Moreover, scant DCM research has directly explored how stakeholders interpret and use DCM-based diagnostic feedback, resulting in some uncertainty about its application in real-world settings (Lee, 2015; Sessoms & Henson, 2018). Nevertheless, informed test score interpretation and utilization are especially critical for high-stakes tests used to determine EAL students' language proficiency for post-secondary admission decisions.

Research on the use of standardized test scores in higher education institutions reveals significant variations in attitudes and perceptions of language proficiency tests among test score users (O'Loughlin, 2008). Studies indicate that many test score users have limited understanding of test score meanings and the rationale behind the admission cut-offs (Coleman et al., 2003; Ockey & Gokturk, 2019). Additionally, insufficient training is available for university administrative and academic staff (Rea-Dickins et al., 2007). Consequently, these users seek more in-depth information on interpreting test scores, especially regarding what students typically understand and can do at different proficiency levels (Hyatt & Brooks, 2006; McDowall & Merrylees, 1998). This limited awareness among test score users about test score interpretations significantly impacts the admission decision-making process and resource allocations for post-admission support (O'Loughlin, 2008). O'Loughlin (2013) emphasizes the importance of test score users having a deep understanding of assessments for ensuring valid practices. He points out that regardless of a language test's technical robustness, the significance and use of its scores ultimately rely on the users of these scores.

The present study investigated the potential of DCM-based reading skill mastery profiling. Specifically, it addressed the following two research questions: 1) What characteristics do reading skill profiles exhibit across the IELTS band scores? and 2) How do test score users interpret these reading skill profiles in relation to admission decisions and support programming?

Method

Rationale for Mixed Methods Research

Brown (2014) argues that mixed methods research should integrate the characteristics of qualitative and quantitative data in such a way that they complement each other, making the mixed methods research greater than the mere sum of its qualitative and quantitative parts (p. 127). In the current study, a two-phase sequential mixed methods research design was applied. In this design, reading skill profiles, estimated through the application of DCMs to large-scale IELTS reading test response data, were used to engage test score users in interpreting the reading skill profiles and planning for their use.

Participants

An item response dataset from the IELTS reading test section, consisting of 5,222 test takers (48.6% female), was used to estimate skill mastery profiles across the IELTS band scores

through the application of DCM. The data files and test materials were provided by Cambridge English through the IELTS® Research Grant Program. Test takers' ages ranged from 14 to 59 years with an average age of 25.45 (*SD* = 6.92). Over 69% of the test takers (*n* = 3,609) took the test for higher education admission. Approximately 8.4% took the test for medical registration, 6% for immigration, 3.4% for employment, and 3.2% for non-medical professional registration. Test takers reported 135 different first languages, with approximately 33% listing Chinese, followed by Tagalog (6.2%), Urdu (5.3%), Arabic (5%), English (4%), and Bengali (3.6%).

We conducted two focus groups to examine how test score users interpret IELTS reading skill profiles derived from the DCM application. The first focus group comprised six undergraduate students from a Canadian research university who had taken the IELTS for admission purposes. The second focus group consisted of six faculty members and administrators from the same university. The faculty participants were affiliated with two different programs, while administrators were representatives from the Registrar's Office, Enrolment Services, and the Office of the Vice-President and Provost.

Measure and Materials

The IELTS Academic, recognized as one of the most widely used English language proficiency tests by English-medium post-secondary institutions worldwide, comprises reading, writing, speaking, and listening sections. The IELTS Academic reading test takes an hour and includes 40 items based on three reading passages about academic topics. Every item response was scored dichotomously: 1 for a correct response and 0 for an incorrect one. The overall internal consistency of the IELTS reading test, as measured by the Kuder-Richardson 20 formula, was .86. The difficulty levels of the items varied: 19 of the 40 items exhibited moderate item difficulty with proportion-correct scores (*p*-value) between .40 to .70, whereas ten items were relatively easier, each with a *p*-value over .7. IELTS reading test scores are calculated by converting raw scores (with a maximum of 40) to the IELTS nine-band scale, and they are reported in whole or half bands (see more details at https://www.ielts.org/for-organisations/ielts-scoring-in-detail). The mean band score was 5.91, with a standard deviation of 1.12. Approximately 22% of the sample received 5.5, followed by 18.3% and 18.2% who scored 5.0 and 6.0, respectively.

After establishing the skill profiles across the IELTS reading band scores using DCM applications, we conducted two focus groups to examine how students, faculty members, and administrative staff interpret the resulting reading skill profiles. The participants offered feedback on these profiles and discussed ways to assist international EAL students.

Data Analysis

The IELTS reading test item response data were analyzed using DCMs with the "CDM" package in the R statistical software (Robitzsch et al., 2020). The CDM package employs an expectation maximization algorithm to estimate skill mastery probabilities based on marginal maximum likelihood estimation. The two primary data sources for DCMs include the item response data and a weighted matrix known as a Q-matrix. This matrix specifies the relationship between items and user-specified skills (Tatsuoka, 1983).

To develop the Q-matrix, seven content experts in applied linguistics and language testing, including two males and three speakers of English as an additional language, identified 11 reading skills relevant to the IELTS reading test. This process involved reviewing literature on reading skill taxonomies and strategies. Item content analysis revealed that most items measured explicit comprehension, inferencing, and summarizing. In contrast, other skills were measured by two or fewer items, necessitating revision due to insufficient number of items for accurate classifications (Hartz et al., 2002; Jang, 2005). The initial Q-matrix included seven skills; however, only three skills had a sufficient number of items (minimum 3-4 items per skill) for reliable DCM estimation (Templin & Bradshaw, 2013). Out of the total 40 items, 26 items were designed to assess explicit text comprehension, 12 focused on inferential reasoning, and 10 tested summarizing skills. The final Q matrix included these three reading skills.

Explicit text comprehension is defined as the basic comprehension of textual information involving processing explicitly stated information from one or two sentences at a local level. Inferential reasoning requires readers to reason beyond the given text, drawing upon background knowledge or textual information to generate hypotheses, predict future events, or infer the author's purpose. Summarizing is defined as the skill required to comprehend key ideas on a global level by connecting, integrating, and summarizing information from multiple sentences or paragraphs, recognizing the text's organizational structure, and distinguishing main ideas from supporting details. On the few instances where discrepancies in item coding based on these three skills arose among the content experts, they reviewed the items, skill definitions, and coding scheme collectively until a unanimous consensus was reached.

DCMs differ by their psychometric parameterizations, estimation algorithms, and theoretical assumptions about inter-skill relationships, such as compensatory versus conjunctive. Therefore, we included five of the most well-known psychometric models (i.e., DINA, DINO, G-DINA, NC-RRUM, C-RRUM) for comparative analyses (e.g., Effatpanah, 2019; Lee & Sawaki, 2009; Ravand & Robitzsch, 2018). We employed various evaluation techniques to assess the fit of each model. In comparing model fits, we considered both sensitivity-having sufficient parameters to model relationships among variables- and specificity to avoid model overfitting. We used several information-based model evaluation criteria for model comparison, including Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), AIC3 (corrected AIC with a penalty factor of 3), and Consistent Akaike's Information Criterion (CAIC). All of these are widely accepted goodness-of-fit criteria with penalties to control for overfitting.

Data from the focus groups were transcribed and subsequently analyzed for key themes (Coleman & O'Connor, 2007). We paid close attention to the focus group participants' interpretations of and feedback on the skill profiles, the challenges international EAL students face in their coursework, and suggestions for future programming to support EAL students upon admission.

Results

Characteristics of Reading Skill Profiles

As shown in Table 1, the AIC, BIC, AIC3, and CAIC goodness-of-fit criteria indicated that the G-DINA model, together with C-RRUM, demonstrated a good model fit for the given response

data. The average Root Mean Squared Error (RMSE) for G-DINA was .047, which is below the recommended cut-off value of .05 (Templin & Henson, 2006). The RMSE quantifies the difference between the predicted and observed correlations for all item pairs (Yi, 2017).

Table 1

M*odel Fit Comparison (N = 5,222)*

Model	No of parameters	loglike	AIC	BIC	AIC3	CAIC
DINA	87	-116792.4	233758.8	234329.6	233845.8	234416.6
DINO	87	-117262.7	234699.4	235270.2	234786.4	235357.2
G-DINA	**109**	-116461.7	**233141.4**	**233856.5**	**233250.4**	**233965.5**
NC-RRUM	98	-117970.1	236136.1	236779.1	236234.1	236877.1
C-RRUM	98	-116485.3	233166.6	233809.6	233264.6	233907.6

Note. DINA (deterministic inputs, noisy "and" gate); DINO (deterministic-input, noisy-or-gate); G-DINA (generalized deterministic inputs, noisy "and" gate); NC-RRUM (non-compensatory reduced reparameterized unified model); C-RRUM (compensatory RRUM).

Table 2 presents the likelihood ratio tests for comparing the nested models. Overall, the G-DINA model showed a better fit compared to the other models, suggesting that correctly answering multi-skill reading items requires an interaction of reading skills. Thus, we used the results from the G-DINA application to address the first research question.

Table 2

Likelihood Ratio Tests for Model Comparison

Test	Model 1	Model 2	$\chi 2$	*df*	*p*
1	DINA	G-DINA	661.40	22	< .01
2	DINA	NC-RRUM	-2355.32	11	> .99
3	DINA	C-RRUM	614.18	11	< .01
4	DINO	G-DINA	1601.98	22	< .01
5	DINO	NC-RRUM	-1414.74	11	> .99
6	DINO	C-RRUM	1554.76	11	< .01
7	NC-RRUM	G-DINA	3016.72	11	< .01
8	C-RRUM	G-DINA	47.22	11	< .01

The G-DINA model parameters and skill patterns were estimated using maximum likelihood estimation (MLE) (Ma & de la Torre, 2019). The results from the G-DINA application showed that most items effectively distinguished skill masters from non-masters. Table 3 presents the final Q-matrix along with item parameter estimates.

Table 3

Item and Attribute Discrimination Values of the Final Q-Matrix

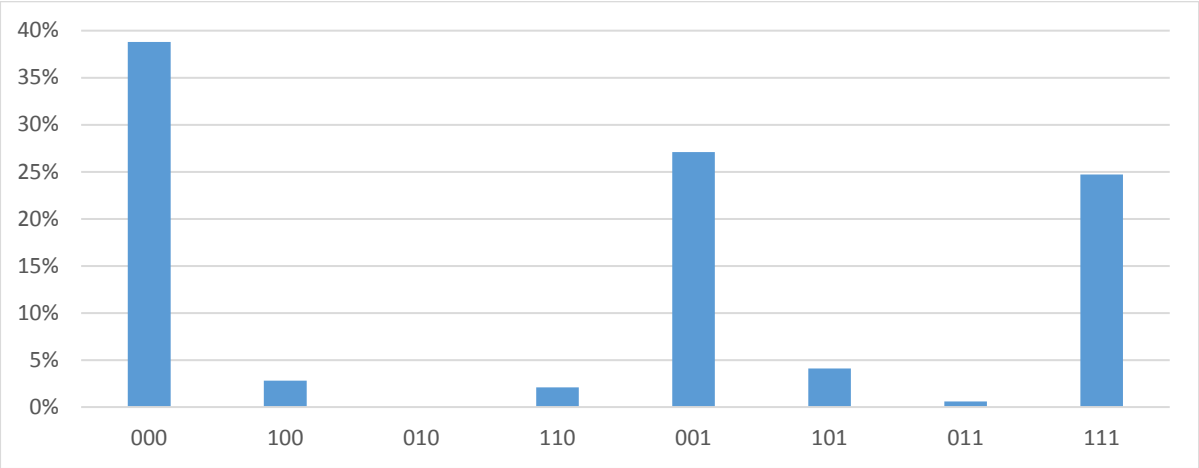
Item	Item Discrimination Value			*p-value*	Global Item Discrimination Value
	Explicit Text Comprehension	Inferential Reasoning	Summarizing		
1			1.82	.71	1.21
2			1.25	.65	0.83
3			0.67	.81	0.45
4			0.27	.86	0.18
5	0.03			.10	0.02
6	0.44			.42	0.29
7		0.33		.68	0.22
8	0.06	0.05		.71	0.05
9	0.21			.37	0.14
10	0.21		0.04	.68	0.17
11	0.21			.50	0.09
12	0.44		0.06	.31	0.33
13	0.32		0.03	.29	0.23
14		0.30		.58	0.20
15	0.24			.69	0.16
16	0.14			.85	0.09
17		0.08		.72	0.06
18	0.01			.87	0.01
19		0.17	0.05	.36	0.12
20	0.08			.74	0.05
21		0.14		.76	0.10
22	0.15			.55	0.10
23	0.21			.47	0.14
24	0.32	1.31		.75	0.63
25	0.15			.71	0.10
26	0.30			.47	0.20
27			0.16	.68	0.11
28		0.13	0.01	.49	0.09
29		0.16		.50	0.11
30		0.01		.35	0.00
31		0.34		.41	0.23
32	0.81			.62	0.54
33	0.39		0.04	.56	0.28
34		0.63		.52	0.42
35	0.61			.52	0.40

36	0.34			.61	0.22
37		1.13	0.02	.22	0.77
38	0.87			.35	0.58
39	1.10	0.08		.27	0.70
40	0.57			.32	0.38
Att. Dis. Value	8.09	4.88	4.43		

Among the three reading skills, the summarizing skill showed the highest proportion of mastery (56.5%), while the explicit text comprehension skill showed 35% skill mastery in the sample. Inferential reasoning skill showed mastery in 28% of the sample, the lowest among all skills. Figure 1 shows the distribution of reading skill mastery profiles based on the application of a .5 cut-off (Kim, 2015; Lee & Sawaki, 2009; Li, 2011; Ravand, 2016; Ravand & Robitzsch, 2015; Yi, 2017) to the posterior probability estimate of skill mastery. Out of the 8 possible mastery classes (equivalent to 2^3), approximately 39% of the sample did not master any skill (000), while 25% mastered all three skills (111).

Figure 1

Frequency of Skill Mastery Profiles



Note. The order of the three skills within each attribute pattern is: Exp-Inf-Sum. Exp: Explicit Text Comprehension. Inf: Inferential Reasoning. Sum: Summarizing. Zero (0) refers to the non-mastery of the given skill. One (1) refers to the mastery of the given skill.

Table 4 shows the distribution of reading skill profiles across the IELTS reading band scores. As shown in Table 4, students with an IELTS band score of 6.5, which is a common cut score used for admission, typically demonstrated mastery in two skills: explicit text comprehension and summarizing. However, their probability of mastering inferential reasoning was significantly lower.

Table 4

Average Posterior Skill Mastery Probability Estimates across the IELTS Band Scores

IELTS Band Scores	Posterior Skill Mastery Probabilities		
	Explicit Text Comprehension	Inferential Reasoning	Summarizing
4.5	0	0	.10
5	0	0	.25
5.5	.03	0	.48
6	.26	.12	.68
6.5	**.81**	**.58**	**.85**
7	1.00	.89	.92
7.5	1.00	.98	.97
8	1.00	.99	.99
8.5	1.00	1.00	1.00
9	1.00	1.00	1.00

Test Score Users' Interpretations of the Skill Profiles

The focus group participants reviewed the skill profiles across the band scores (Table 5) and shared their insights on its interpretation and usefulness. Both students and faculty members believed this information would be invaluable if included in the test score report or alongside the university's acceptance letter. After reviewing the can-do descriptors, students recognized the challenges they faced transitioning from high school to university, particularly in the shift in academic language expectations. Language demands varied across fields of study. Some required understanding of statistical tables, figures, or formulas, which were particularly challenging for those unfamiliar with the context of their target language. Faculty members were surprised at the descriptors for the cut-off score of 6.5, having assumed students at this level could handle the academic demands associated with scores of 7.0 or higher. All participants highlighted the importance of critical thinking and inferential reasoning, which the skill profiles indicated were lacking for those scoring 6.5. They noted that most courses demand engagement with extensive reading materials and critical appraisal of diverse viewpoints.

Table 5

Can-Do Skill Profiles for IELTS Reading Band Scores

Band Score	Students can
5.5	• Locate a keyword or a topic sentence by scanning and skimming a text. • Comprehend the literal meaning of short phrases or simple sentences. • Figure out the meaning of high-frequency vocabulary.
6.0	• Understand the main idea from a paragraph. • Distinguish the main idea from supporting details. • Figure out the meaning of moderately difficult vocabulary.
6.5	• Comprehend the implicit meaning in the text. • Paraphrase main ideas. • Summarize the main idea from a long, grammatically complex text. • Figure out the meaning of low-frequency vocabulary. • Begin to infer implicit meaning from the text.

7.0	• Synthesize the main idea with supporting details from the text. • Make inferences about implicit information from the text. • Understand the logical connections between ideas across sentences.
7.5	• Infer meaning in the text specific to a certain culture. • Figure out colloquial expressions in the text. • Comprehend text that contains abstract vocabulary and features grammatically complex sentence structures (e.g., if-then, although-)

Students discussed the challenges they faced when transitioning from high school to university, including understanding culturally specific meanings, collocations, and idiomatic expressions. International EAL students, in particular, may face challenges stemming from cultural knowledge gaps. The students suggested more detailed post-admission support, actionable steps to develop skills, samples of academic texts, and real-world academic tasks. They also sought examples connecting the 'can-do' reading skill profiles to academic requirements and resources for skill improvement. Faculty and staff voiced concerns about international students' readiness upon admission and their sense of isolation due to language barriers. The discussion touched on the balance between attractive admission requirements and ensuring student success. They extensively discussed support strategies for incoming students, considering the timing and communication modes suitable for international EAL students.

Discussion

DCM approaches have had limited applications to real-life testing practices due to various factors, including a lack of tests developed with DCM design principles and the computational intensity of estimation methods (von Davier, 2008). Given that current testing practice heavily relies on unidimensional scaling, recent research shows some promise in blending DCM with other unidimensional techniques (Bradshaw & Templin, 2014; Choi, 2010; Tseng & Wang, 2021). For example, Choi (2010) incorporated skill mastery states as covariates in combining a log-linear DCM with a Mixture Rasch model. Tseng and Wang (2021) examined the potential of the Q-matrix anchored mixture Rasch model by using class invariant items (Paek & Cho, 2015; von Davier & Yamamoto, 2004) and aligning model parameter estimates from different latent classes on a common scale. Future research should examine possibilities for providing the mastery states of skills, as identified by content experts, while also providing an overall language ability score to guide the decision-making processes.

The results from DCM analysis indicate that students meeting the institutional cut-off score for admission (e.g., 6.5 in this study's context) may not fully master inferential reasoning. For example, the average posterior probability estimate for mastering inferential skills at the 6.5 IELTS band score was slightly over .5. This aligns with the IELTS's (2014) guidance for educational institutions that a 6.5 overall score is likely acceptable only for less linguistically demanding programs or courses. Considering that academic tasks require critical and global reading of extensive texts, students with an IELTS band score of 6.5 may struggle with academic tasks requiring advanced reading skills, such as inferential reasoning.

Similar to previous research (Ginther & Elder, 2014; Green et al., 2010; Weir et al., 2009), our study findings show that most IELTS reading test items tend to focus on explicit textual

comprehension at the sentence or local level. It has been recommended that the IELTS reading test specifications include more items related to global reading as well as tasks that prompt test takers to evaluate multiple sources, as students often struggle with real-life academic reading demands involving extensive reading under tight time constraints (Bax, 2015; Moore et al., 2012). Other DCM research on IELTS, such as studies by Aryadoust (2012), Effatpanah (2019), and Mirzaei et al. (2020), reports similar findings. For example, Effatpanah (2019) applied the G-DINA model to the IELTS listening test and reported that making inferences was particularly challenging for Iranian test takers. Aryadoust (2012) reported that most items tend to tap into basic comprehension skills, which raises concerns about the test's underrepresentation of higher-order skills.

Previous research on the academic achievements of EAL students in the context of English for academic purposes (Dang & Dang, 2021; Hamp-Lyons, 2011; Rajendram et al., 2019; Sawir et al., 2012), discipline-specific literacy practices in academic settings (Shanahan & Shanahan, 2012), and test takers' experiences with post-admission language demands (Clark & Yu, 2021; Pearson, 2020), suggests that many students admitted to universities often lack higher-order skills. Merely raising cut-off scores for admission does not solve these challenges since increasing cut-off scores does not necessarily bridge the gap between real-life academic language demands and the competencies standardized tests measure (MacDonald, 2019; O'Loughlin, 2011).

In the present study, both students and faculty members participating in focus groups expressed appreciation for detailed information about test scores, specifically highlighting what students can and cannot do at different score levels. When provided with the "can-do" reading skill profiles associated with the IELTS reading band scores, test score users engaged in discussions about ways to improve and support student needs. Both test developers and test score users (e.g., post-secondary institutions) should actively collaborate, taking on shared responsibilities. There is also a compelling need to promote test literacy among test score users (Baker, 2016; Ockey & Gokturk, 2019). By doing so, institutions can better support faculty members in meeting student needs, ensuring that students are better prepared and supported to meet the language demands of their coursework.

Our study is limited in overcoming such issues as it retrofitted DCMs to the IELTS reading test response data. The initial Q-matrix we developed with content experts included a more comprehensive range of reading skills, better aligned with reading theories and reading tasks in academic settings. However, the final Q-matrix used for DCMs included only three skills. Furthermore, the retrofitted DCMs inevitably resulted in an uneven number of items per skill for DCM applications. If a given test does not have enough items for each skill, determining whether test takers have mastered it becomes unreliable. We concur with Bejar et al. (2007) in asserting that skill descriptors should be explicitly incorporated as part of a test design and item specifications. Future research should address these limitations, ideally by incorporating theoretically defensible skill specifications into item design rather than by retrofitting.

Conclusion

In the contemporary perspective of test validity, ensuring meaningful test score interpretations and appropriate use of test scores is critical. Both test developers and test score users have a

shared commitment to promoting valid, reliable, and fair assessment practices (Chalhoub-Deville & Turner, 2000). Test score users are responsible for appropriate test use, and they should carry out local investigations to ensure that their admission requirements align with the academic demands required for success in specific programs. Conversely, test developers should be responsible for developing tests that meet established professional standards, providing high-quality information about students' abilities, and providing user guides to support decision-making processes in local contexts (AERA, NCME, & APA, 2014; O'Loughlin, 2013).

Despite the potential benefits of DCM, its utilization in real-life test score reporting practice remains limited, partly due to the stringent theoretical and psychometric demands of DCM models as scaling alternatives (Rupp & Templin, 2008). DCM's formative, diagnostic potential through skill profiling may be realized when tests used for high-stake decisions are carefully designed and calibrated with DCM specifications.

Design and validation efforts centered around DCMs should actively involve the end-users of test scores. Such collaboration ensures the creation of diagnostic profiles that not only provide interpretable insights but also guide relevant actions. Moreover, there is a strong need for future research to delve deeper into the effects of DCM-based diagnostic feedback. Specifically, future research should examine how this feedback influences EAL students' readiness for academic pursuits, as well as the long-term progression of their language proficiency.

In conducting such future research, we echo Brown's (2014) call for embracing a culture of mixed methods research. This approach allows us to recognize multiple research avenues from diverse possibilities and to "consolidate them through selected forms of legitimation into research that is more than reliable or dependable, more than valid or credible, more than replicable or confirmable, and more than generalizable or transferable" (p. 232).

ORCID

https://orcid.org/0000-0002-8824-8216

https://orcid.org/0000-0002-5883-7714

https://orcid.org/0000-0002-8607-9033

https://orcid.org/0009-0000-6297-6509

Acknowledgements

This study was funded by the Cambridge English Language Assessment IELTS® Research Grant Program (Grant Number: AU2212). We are grateful for the data files, test materials, and the constructive feedback provided on our study report.

Funding

Cambridge English Language Assessment IELTS® Research Grant Program (Grant Number: AU2212).

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Aryadoust, V. (2012). Differential item functioning in while-listening performance tests: The case of the International English Language Testing System (IELTS) listening module. *International Journal of Listening*, *26*(1), 40-60. https://doi.org/10.1080/10904018.2012.639649

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Baker, B. (2016). Language assessment literacy as professional competence: The case of Canadian admissions decision makers. *Canadian Journal of Applied Linguistics*, *19*(1), 63–83. https://journals.lib.unb.ca/index.php/CJAL/article/view/23033

Bax, S. (2015). Using eye-tracking to research the cognitive processes of multinational readers during an IELTS reading test. *IELTS Research Reports Online Series*, 21.

Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, predictive, and progressive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (pp. 1-30). Maple Grove, MN: Journal of Applied Metrics Press.

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: a psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika, 79*, 403-425. https://doi.org/10.1007/s11336-013-9350-4

Brown, J. (2014). *Mixed methods research for TESOL*. Edinburgh University Press.

Chalhoub-Deville, M., & Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System*, *28*(4), 523–539. https://doi.org/10.1016/S0346-251X(00)00036-1

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Choi, H. J. (2010). *A model that combines diagnostic classification assessment with mixture item response theory models*. Georgia: University of Georgia.

Coleman, D., Starfield, S., & Hagan, A. (2003). *The attitude of IELTS stakeholders: student and staff perceptions of IELTS in Australia, UK and Chinese tertiary institutions in IELTS*. Research Reports (V.5). IELTS Australia, Canberra.

Coleman, G., & O'Connor, R. (2007). Using grounded theory to understand software process improvement: A study of Irish software companies. *Information and Software Technology*, 49(6), 654–667. https://doi.org/10.1016/j.infsof.2007.02.011

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

Clark, T. & Yu, G. (2021). Beyond the IELTS test: Chinese and Japanese postgraduate UK experiences, *International Journal of Bilingual Education and Bilingualism*, *24*(10), 1512–1530. https://doi.org/10.1080/13670050.2020.1829538

Dang, C., & Dang, T. (2021). The predictive validity of the IELTS test and contribution of IELTS preparation courses to international students' subsequent academic study: Insights from Vietnamese international students in the UK. *RELC Journal, 54*(1), 84-98. https://doi.org/10.1177/0033688220985533

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

DiBello, L. V., & Stout, W. (2007). Guest Editors' Introduction and Overview: IRT-Based Cognitive Diagnostic Models and Related Methods. *Journal of Educational Measurement*, *44*(4), 285–291. https://www.jstor.org/stable/20461864

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols & S. F. Chipman, & Brennan, R. L. (Eds.), *Cognitively diagnostic assessment* (pp. *361-389)*. Routledge.

Dong, Y., Ma, X., Wang, C., & 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

Effatpanah, F. (2019). Application of diagnostic classification models to the listening section of the International English Language Testing System (IELTS). *International Journal of Language Testing, 9*(1), 1-28. https://files.eric.ed.gov/fulltext/EJ1299308.pdf

Ginther, A., & Elder, C. (2014). A comparative investigation into understandings and uses of the TOEFL iBT® Test, the International English Language Testing Service (Academic) test, and the Pearson Test of English for graduate admissions in the United States and Australia: A case study of two university contexts. *ETS Research Report Series*, *2014*(2), 1–39. https://doi.org/10.1002/ets2.12037

Green, A., Ünaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing, 27*(2), 191–211. https://doi.org/10.1177/0265532209349471

Hamp-Lyons, L. (2011). English for academic purposes. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning (*Vol. *2*, pp. 89-105). New York, NY: Routledge.

Hartz, S., Roussos, L., & Stout, W. (2002). *Skills diagnosis: Theory and practice. User manual for Arpeggio software*. ETS.

Hyatt, D., & Brooks, G. (2006). Investigating stakeholders' perceptions of IELTS as an entry requirement for higher education in the UK. *IELTS Research Reports*, *10*(1). 3–50.

IELTS. (2014). *Guide for educational institutions, governments, professional bodies and commercial organisations*. IELTS.

Ingram, D., & Bayliss, A. (2007). IELTS as a predictor of academic language performance. *IELTS Research Reports*, *7*, 137–204.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. University of Illinois at Urbana-Champaign.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing*, *26*(1), 31-73. https://doi.org/10.1177/0265532208097336

Jang, E. E., Dunlop, M., Park, G., & Van Der Boom, E. H. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Language Testing*, *32*(3), 359–383. https://doi.org/10.1177/0265532215570924

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

Kane, M. T. (2006). Validation. In R. B. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger. https://doi.org/10.1037/0033-2909.112.3.527

Kim, A. Y. (2015). 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

Lee, Y., & Sawaki, Y. (2009). 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. (2015). Diagnosing diagnostic language assessment. *Language Testing*, *32*(3), 299–316. https://doi.org/10.1177/0265532214565387

Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Spaan Fellow*, *9*, 17-46.

Li, H., Hunter, C. V., & Lei, P. W. (2016). The selection of diagnostic classification models for a reading comprehension test. *Language Testing, 33*(3), 391-409. https://doi.org/10.1080/0969594x.2015.1060192

MacDonald, J. (2019). Sitting at 6.5: Problematizing IELTS and admissions to Canadian universities. *TESL Canada Journal, 36*(1), 160-171. https://doi.org/10.18806/tesl.v36i1.1308

McDowall, C., & Merrylees, B. (1998). Survey of receiving institutions' use and attitude to IELTS. In S. Wood (Ed.), *International English Language Testing System (IELTS) Research Reports (*Vo. 1, pp. 116-139). Canberra: IELTS Australia.

Milanovic, M., & Weir, C. J. (2010). Series editors' note. *Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's Draft Manual*, viii–xx.

Min, S., & He, L. (2022). Developing individualized feedback for listening assessment: Combining standard setting and cognitive diagnostic assessment approaches. *Language Testing*, *39*(1), 90–116. https://doi.org/10.1177/0265532221995475

Min, S., Cai, H., & He, L. (2022). Application of Bi-factor MIRT and Higher-order CDM Models to an In-house EFL Listening Test for Diagnostic Purposes. *Language Assessment Quarterly*, *19*(2), 189–213. https://doi.org/10.1080/15434303.2021.1980571

Mirzaei, A., Heidari Vincheh, M., & Hashemian, M. (2020). Retrofitting the IELTS reading section with a general diagnostic classification model in an Iranian EAP context. *Studies in Educational Evaluation*, *64*, 100817. https://doi.org/10.1016/j.stueduc.2019.100817

Moore, T., Morton, J., & Price, S. (2012). Construct validity in the IELTS Academic Reading test: A comparison of reading requirements in IELTS test items and in university study. *IELTS Research Reports, 11*(4), 1–86. https://www.ielts.org/-/media/research-reports/ielts_rr_volume11_report4.ashx

Ockey, G. J., & Gokturk, N. (2019). Standardized language proficiency tests in higher education. In X. Gao (ed.), *Second handbook of English language teaching* (pp. 377-393), Springer. https://doi.org/10.1007/978-3-030-02899-2_25

O'Loughlin, K. (2008). The use of IELTS for university selection in Australia: A case study. *IELTS Research Reports*, *8*(3), 3–98.

O'Loughlin, K. (2011). The interpretation and use of proficiency test scores in university selection: How valid and ethical are they? *Language Assessment Quarterly*, *8*(2), 146-160. https://doi.org/10.1080/15434303.2011.564698

O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing*, *30*(3), 363-380. https://doi.org/10.1177/0265532213480336

Paek, I., & Cho, S. J. (2015). A note on parameter estimate comparability across latent classes in mixture IRT modeling. *Applied Psychology Measurement, 39*, 135-143. https://doi.org/10.1177/0146621614549651

Pearson, W. S. (2020). Mapping English language proficiency cut-off scores and pre-sessional EAP programmes in UK higher education. *Journal of English for Academic Purposes*, *45*, 100866. https://doi.org/10.1016/j.jeap.2020.100866

Rajendram, S., Sinclair, J., & Larson, E. (2019). International graduate students' perspectives on high-stakes English tests and the language demands of higher education. *Language & Literacy*, *21*(4), 68–92. https://doi.org/10.20360/langandlit29428

Ravand, H. (2015). Assessing testlet effect, impact, differential testlet, and item functioning using cross-classified multilevel measurement modeling. *SAGE Open*, *5*(2). https://doi.org/10.1177/2158244015585607

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

Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research, and Evaluation*, *20*, 1-12. https://doi.org/10.7275/5g6f-ak15

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

Rea-Dickins, P., Kiely, R., & Yu, G. (2007). Student identity, learning and progression: The affective and academic impact of IELTS on 'successful' candidates. *IELTS Research Reports, 7*(2), 2–78.

Robitzsch, A., Kiefer, T., George, A. C., & Ünlü, A. (2020). *DCM: Cognitive Diagnosis Modeling. R package version 7.5-15*. https://CRAN.R-project.org/package=DCM

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

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

Sawir, E., Marginson, S., Forbes-Mewett, H., Nyland, C., & Ramia, G. (2012). International student security and English language proficiency. *Journal of Studies in International Education, 16*(5), 434-454. https://doi.org/10.1177/1028315311435418

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 1-17. https://doi.org/10.1080/15366367.2018.1435104

Shanahan, T., & Shanahan, C. (2012). What is disciplinary literacy and why does it matter? *Topics in Language Disorders*, *32*(1), 7-18. DOI: 10.1097/TLD.0b013e318244557a

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 345-354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Taylor, L., & Weir, C. J. (Eds.). (2012). *IELTS collected papers 2: Research in reading and listening assessment*. Studies in Language Testing 34. Cambridge University Press.

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

Toprak, T. E., & Cakir, A. (2021). Examining the L2 reading comprehension ability of adult ELLs: Developing a diagnostic test within the cognitive diagnostic assessment framework. *Language Testing*, *38*(1), 106-131. https://doi.org/10.1177/0265532220941470

Tseng, M. C., & Wang, W. C. (2021). The Q-matrix anchored mixture Rasch model. *Frontier of Psychology, 12*, 564976. https://doi.org/10.3389/fpsyg.2021.564976

von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series*, *2005*(2), i-35. https://doi.org/10.1002/j.2333-8504.2005.tb01993.x

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

von Davier, M., & Lee, Y. S. (2019). Introduction: From latent classes to cognitive diagnostic models: In von Davier, M., & Lee, Y. S. (Eds.), *Handbook of diagnostic classification models, Methodology of educational measurement and assessment* (pp. 1-17). Springer. https://doi.org/10.1007/978-3-030-05584-4_1

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: an extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*, 389-406. https://doi.org/10.1177/0146621604268734

Weir, C., Hawkey, R., Green, A., Unaldi, A., & Devi, S. (2009). The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university. *IELTS Research Reports*, *9*(3), 97–156.

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

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

Zhai, X., Haudek, K. C., & Ma, W. (2022). Assessing argumentation using machine learning and cognitive diagnostic modeling. *Research in Science Education, 53, 405-424*. https://doi.org/10.1007/s11165-022-10062-w