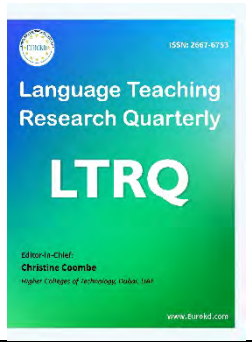

Language Teaching Research Quarterly

2022, Vol. 29, 147–160

Cognitive Diagnostic Assessment of IELTS Listening: Providing Feedback from its Internal Structure

Ali Panahi[1]*, Hassan Mohebbi[2]

[1]Iranian English Language Institute, Ardebil, Iran

[2]European Knowledge Development Institute, Turkey

Received *07 February 2022* ***Accepted*** *28 April 2022*

Abstract

High stakes testing, such as IELTS, is designed to select individuals for decision-making purposes (Fulcher, 2013b). Hence, there is a slow-growing stream of research investigating the subskills of IELTS listening and, in feedback terms, its effects on individuals and educational programs. Here, cognitive diagnostic assessment (CDA) performs it through breaking down the abstract listening concepts into fine-grained subcomponents, and applying it to in-class assessment and teaching. Therefore, the strengths and weaknesses of language learners concerning various subsections of the IELTS listening test were explored with use of three CDMs, i.e., DINA, DINO and mixed DINA-DINO. As a result, the analysis of the participants'(N=463) performances revealed that the mixed-DINA-DINO model using the original Q-matrix was the most effective model. On the other hand, all three models indicated that gap-filling was cognitively less demanding than the other subsections. However, they did not show any agreement on multiple-choice sub-competency. Accordingly, it can be argued that the CDM-driven diagnostic information related to the sub-sections of the IELTS listening test can be used in educational systems to explore the underlying structure of a test and detect the learners' potential strengths and weaknesses with mastery and non-mastery of the intended items.

Keywords: *Cognitive Diagnostic Assessment, CDA, IELTS Listening Test*

Introduction

In the first decade of the 21st century, following a staggering growth in language testing and assessment, heavy actions and responsibilities have been placed upon language teachers (Fulcher,

hafez1358720@gmail.com

doi:10.32038/ltrq.2022.29.10

2012). Assessment provides feedback for the teachers and testers and this renders it demanding to find out and present appropriate feedback. Clearly, feedback is one of the crucial factors affecting the learners' success and achievement (Hattie & Timperley, 2007), so students urgently need to experience the value of feedback (Carless & Boud, 2018; Leighton, 2019). Admitting the significant contribution of feedback to both education and individuals, educators need statistical tools to uncover the hidden layers of learners' mastery and non-mystery of the intended content. A variety of sources are used to develop the learning and learning strategies of the students (Carless & Boud, 2018) as these strategies and tools can help diagnose the potential strengths and weaknesses of learners or test takers. One of these tools is cognitive diagnostic assessment (CDA), as it offers more effective feedback regarding test takers or learners' mastery or non-mastery of specific subskills (Aryadoust, 2019, 2018; DeCarlo, 2010; Min & He, 2021). Regarding the way CDA is used to analyze the items, Ravand and Robitzsch (2018) stated that CDM analyses items in terms of the real performance of examinees on the sub-skills or item-types; this capacitates the teachers to know the skills and sub-skills the students have mastered or they need to improve.

Listening facilitates and contributes to interaction and development of other language skills (Bodie & Worthington, 2010; Vandergrift & Goh, 2012), some dearth of research is, however, observed (Badger & Yan, 2006; Aryadoust, 2012; Phakiti, 2016) in connection with CDA-related IELTS listening test research. In particular, very little research has so far been conducted into CDM-relevant feedback on listening. Recently, Min and He (2021) investigated the development of individualized feedback for listening using CDA approaches. They used a large sample size of 3358 EFL students. Their findings indicated that CDA model facilitates the way teachers direct and apply feedback. Therefore, the primary purpose of CDA in education and language assessment is to diagnose and assess examinees' mastery and non-mastery of skills or attributes (Chen et al., 2013; De La Torre, 2011; George & Robitzsch, 2015; Rupp & Templin, 2008).

CDM is a diagnostic system (Harding et al., 2015; Min & He, 2021) that assists educators and language assessors to bring standardized testing, such as IELTS, into the language learning process, i.e., IELTS preparation course with skills-level examinees' parameters for providing information about the diagnostic power of an intended test; this system is a promising and effective tool for diagnosing the skills and the mastery and non-mastery of the candidates with a particular attribute (Ravand & Robitzsch, 2018).

The context of running CDM is the listening skill of the IELTS test. IELTS has been used since 1989 (Charge & Taylor, 1997) and focuses on language skills in the social and academic context (Nakatsuhara et al., 2017; Phakiti, 2016). In this vein, Pilcher and Richards (2017) indicate the widespread role of IELTS in education and its impact on educational stakeholders. In effect, there is a need for educators in and across a range of IELTS preparation courses to diagnose the weaknesses of language learners with the use of diagnostic tools and attempt to improve them.

A look forward to the cost of the IELTS exam indicates that to move from learning and teaching to the product of the IELTS preparation course, i.e., passing the exam, for instance, the learners are required to incur a considerable amount of financial burden (Hamid, 2016; Pilcher, & Richards, 2017). However, diagnostic information in the fashion of feedback can inform learners and teachers when a particular IELTS candidate can take the test.

The study

For running CDM, the first step is to develop and group the subskills or attributes and then to develop a Q-matrix for the constructs or attributes. A Q-matrix indicates "the requirement for sub-skills in successfully answering each individual item" (Li & Suen, 2013). Regarding the Q-matrix specification, DeCarlo (2012) states that the true Q-matrix is unknown due to the imperfect comprehension of the cognitive processes existing in taking any test. Creating a Q-matrix is sometimes subjective (de la Torre, 2008; Rupp & Templin, 2008). Therefore, in the present study, the Q-matrix was subjectively specified; of course, the perspectives of IELTS instructors and a reference to the design of the IELTS listening test were used to guide this. However, the Q-matrix specification is complex, a look back at the literature can help develop a required Q-matrix (e.g., De La Torre, 2008, 2011; Lei & Li, 2016). The reason for the complexity of the Q-matrix is indicated by some researchers (e.g., Kunina-Habenicht et al., 2012; Rupp & Templin, 2008), pointing out that there is no absolute clear-cut rule for Q-matrix specification, as it can also be primarily subjective. The underspecified Q- matrix is highly acceptable when the sample size is much smaller (Lei & Li, 2016).

A great deal of research has been conducted on cognitive diagnosis of students' learning (e.g., De La Torre, 2011; De La Torre & Douglas, 2004; Junker & Sijtsma, 2001; Li & Wang, 2015), but fewer CDM-driven studies have so far been carried out on the IELTS listening test (Aryadoust, 2018). A very recent study has been done on the listening test of the Singapore-Cambridge General Certificate of education O-level using DINA, DINO, G-DINA, HO-DINA, and R-RUM (Aryadoust, 2018).

In the context of the present study, three models of DINA (deterministic-inputs, noisy, "and" gate), DINO (deterministic-inputs, noisy, "or" gate) and mixed DINA-DINO (Junker & Sijtsma, 2001) were explored. These three models were used to analyze the underlying structure of the IELTS listening test. Running on individual items or a whole test, their effects can be investigated in feedback terms, too. Of course, as Ravand and Robitzsch (2018) indicate, the choice of CDM should be conducted at the item level; in contrast, Sorrel et al. (2017) stated that it could also apply to a whole test. Accordingly, following them, we analyzed the sub-construct level aiming at individual items and assessed four sub-sections of the IELTS listening, i.e., gap-filling, diagram labelling, multiple-choice and short answer using three CDMs, i.e., DINA, DINO and mixed DINA-DINO models. The following research questions are investigated, noting that the mastery and non-mastery of the items can be investigated considering the underlying structure and the upcoming feedback relevant to the process of test design and development.

RQ_1: What are the IELTS listening test items' fit indices using DINA, DINO, and mixed DINA-DINO models?

RQ_2: What are the skill probabilities of four sub-competencies of the IELTS listening test using DINA, DINO, and mixed DINA-DINO models?

RQ_3: What are the skill pattern probabilities of the IELTS listening test using DINA, DINO, and mixed DINA- DINO models?

RQ_4: Which model is preferred over other models regarding the original and Validated Q-matrix?

Method

Participants and Data

The study was conducted in an IELTS mock-test situation at two English language institutes in Iran. The participants (N=480) on IELTS preparation courses took a proficiency test, and then 17 participants' performances were excluded due to extraneous variables, such as much lower performances, and missing some subsections. The final group of participants (n = 463) were bilinguals speaking Azeri-Turkish as their native language and Persian as their official and second language. They were characterized by the same cultural, societal, and educational traits. It is also worth noting that the nature of the data of the present study is based on the data of the first author's PhD dissertation.

Instrument

The IELTS listening tests were extracted from Cambridge IELTS books (Cambridge IELTS, 2016, 2017). Two IELTS listening tests, each including 40 items, were employed in the present study. One was used as a proficiency test because to take the IELTS listening test, the participants first needed to be more proficient at listening rather than being at an elementary level and then they took the main test; the reason why the main test was administered was that after administering the proficiency test, those providing faulty replies and losing some subsections of the test due to a lack of proficiency were removed from the study; the tests included four item types, i.e., gap filling, diagram labelling, multiple-choice and short answer. Regarding the IELTS listening test length, the listening passages were played in 30 minutes, and the participants were given 10 minutes to transfer their answers to the answer sheet. In the end, the collected data were analyzed using the CDM package, i.e., R statistical software version 4.1 (George et al., 2016; Robitzsch et al., 2014).

Results

Table 1 displays the original and the validated Q-matrices for the 40–item IELTS LCT. To validate the matrix, first the attributes, i.e., GF, MC, DL and SA, were specified. Then, based on the answers the participants presented to the prompts and to measure four item types of Gap-Filling (GF), Multiple Choice (MC), Diagram Labelling (DL) and Short Answer (SA), through using R-Package, the following, i.e., Table 1, appeared.

Table 1

Original and Validated Q-Matrices

	Original				Validated			
Items	GF	MC	DL	SA	V1	V2	V3	V4
1	1	0	0	0	1	0	0	1
2	1	0	0	0	1	1	1	0
3	1	0	0	0	1	0	1	1
4	1	0	0	0	1	0	0	0
5	1	0	0	0	1	0	1	0
6	1	0	0	0	1	1	0	0
7	1	0	0	0	1	0	0	0

8	1	0	0	0	1	0	0	0
9	1	0	0	0	1	0	1	1
10	1	0	0	0	1	1	0	0
11	1	0	0	0	1	1	1	1
12	1	0	0	0	1	0	1	1
13	1	0	0	0	1	1	1	0
14	1	0	0	0	1	1	1	0
15	0	1	0	0	0	1	1	1
16	0	1	0	0	0	1	1	0
17	0	1	0	0	0	1	1	0
18	0	1	0	0	0	1	0	0
19	0	1	0	0	0	1	0	1
20	0	1	0	0	0	1	0	0
21	0	0	1	0	1	0	1	0
22	0	0	1	0	1	1	1	1
23	0	0	1	0	1	1	1	0
24	0	0	1	0	1	0	1	0
25	0	0	1	0	0	1	1	0
26	0	0	1	0	0	0	1	0
27	0	0	1	0	0	0	1	0
28	0	0	1	0	0	1	1	0
29	0	0	1	0	0	1	1	0
30	0	0	1	0	0	1	1	1
31	0	0	0	1	1	0	1	1
32	0	0	0	1	0	1	1	1
33	0	0	0	1	0	1	0	1
34	0	0	0	1	0	1	1	1
35	0	0	0	1	1	1	1	1
36	0	0	0	1	1	0	1	1
37	0	0	0	1	1	0	1	1
38	0	0	0	1	1	0	0	1
39	0	0	0	1	1	0	1	1
40	0	0	0	1	0	1	1	1

As it is clear from Table 1, to select any CDMs, some criteria should be considered (Rupp & Templin, 2008): the “observed response variables” and “latent predictor variables” should be measured on a dichotomous scale; i.e. correct = 1, false = 0 and the compensatory and non-compensatory nature of models should also be considered. In the study, the DINA, DINO and mixed DINA-DINO models were selected because, based on the classification presented by Rupp and Templin (2008), all three models required dichotomous data; DINA and DINO were classified as compensatory and non-compensatory models. Some researchers (e.g., Lei, & Li, 2016; Rupp, & Templin, 2008; Wang & Gierl, 2011) emphasize that in a compensatory model, mastery of some

skills will compensate for the non-mastery of other skills, while in non-compensatory models, the lack of mastery of one or more skills is not compensated by other skills; the implication here is that if the test takers answer particular items correctly, it is also probable to guess the items they could not answer correctly and vice versa.

Table 2 displays the results of the absolute and relative fit indices. The mixed DINA-DINO model using the original Q-matrix was the best model. Its Akaike's Information Criterion standing for AIC (23154.80) and Bayesian Information Criterion standing for BIC (23547.88) were the lowest among all models. The mixed model based on the validated matrix was the second-best, and the DINO model based on the validated matrix was the third-best. The DINO model based on the original model was the worst one. Here, by implication, it needs to be mentioned that the models which are considered better or the best, can be relatively not absolutely reliable in specifying the number of items mastered and items non-mastered. Another hidden implication can be the fact that in specifying the strengths and weaknesses of test takers, they can be relatively, rather than absolutely, more efficient.

Considering the absolute fit index of chi-square (MAX2), the original Q-matrix's mixed model enjoyed the lowest MAX2 statistic (20.55, p = .005). This was followed by the DINA and DINO models using the original Q-matrix.

Table 2

Fit Indices of Models

Q-matrix	Model	#Npar	MAX2	p-value	AIC	BIC	MADcor	MADQ3	SRMSR	Mean Item RMSEA
Original	DINA	80	22.04	.002	23178.66	23571.74	.049	.051	.061	.053
	DINO	80	22.04	.002	23178.66	23571.74	.049	.051	.061	.053
	Mixed	80	20.55	.005	23154.80	23547.88	.049	.052	.061	.060
Validated	DINA	80	23.71	.001	23175.86	23568.95	.049	.052	.062	.041
	DINO	80	23.90	.001	23166.76	23599.84	.049	.052	.062	.053
	Mixed	80	23.91	.001	23165.66	23558.75	.049	.051	.062	.052

The MADcor indices were all equal to .049. The other two fit indices of MADq3 and SRMSR were slightly higher than .05. All six models showed mean item fit RMSEA's equal to or less than .60. It seems to be an acceptable level of fit for items, although the DINA model using the validated matrix showed the best fit of .041.

Table 3 displays the skill probabilities, i.e., the probability that the skill was mastered. All three models indicated that 90.5 per cent of gap-filling skill was mastered. However, they did not show agreement on the multiple-choice skill. While the mixed method indicated 91.3% mastery of this skill, the percentages for the DINA and DINO were 64.8. All three methods showed acceptable

agreement on diagram labelling (87 %), although the mixed method showed a slightly higher mastery of this skill. The same was true with the short answer; the mixed method showed a slightly higher mastery of this skill. The implication is that various levels of performances were observed on various subcomponents of IELTS, so extensive practice on numerous number of such attributes and subskills ae required in IELTS preparation courses.

Table 3

Skill Probabilities

	Original			Validated		
	DINA	DINO	Mixed	DINA	DINO	Mixed
Gap Filling	0.905	0.905	0.905	0.902	0.864	0.868
Multiple Choice	0.648	0.648	0.913	0.927	0.716	0.731
Diagram Labeling	0.870	0.870	0.872	0.925	0.834	0.844
Short Answer	0.886	0.886	0.910	0.952	0.525	0.525

The skill probabilities based on the validated matrix showed different patterns; whereas mixed-method showed the highest probability of mastery using the original Q-matrix; the mixed method showed the highest mastery based on the validated Q-matrix. The probabilities for multiple choice (92.7 %), diagram labelling (92.5 %) and short answer (95.2 %) were the highest probabilities under both methods.

Table 4 displays the skill pattern probabilities, which show the percentages of mastery of combinations of skills. The number of skill patterns for any dichotomously rated test can be computed as 2^a, where a stands for the number of skills. Since there are four skills in this study, there will be 16 skill patterns. The results showed that the probabilities for non-mastery of all four skills were 8.94% under DINA and DINO models, while it was 5.77% using a mixed model. The results for the validated Q-matrix showed the opposite pattern. The DINA model had the lowest probability of 2.58%, while the probabilities for the DINO and mixed models were 8.57 % and 8.48 %.

Table 4

Skill Pattern Probabilities

Patterns	Original			Validated		
	DINA	DINO	Mixed	DINA	DINO	Mixed
0000	0.08944	0.08944	0.05776	0.02581	0.08576	0.08485
1000	0.00978	0.00978	0.00876	0.00003	0.01502	0.01516
0100	0.00000	0.00000	0.02208	0.00266	0.00000	0.00000
0010	0.00000	0.00000	0.00000	0.01094	0.00671	0.00647
0001	0.00000	0.00000	0.00788	0.02581	0.00000	0.00000
1100	0.00000	0.00000	0.00000	0.00895	0.03250	0.02790
1010	0.00000	0.00000	0.00000	0.00000	0.01706	0.00850
1001	0.00000	0.00000	0.00000	0.00000	0.00001	0.00000

0110	0.00591	0.00591	0.00017	0.00000	0.00002	0.00000
0101	0.00000	0.00000	0.00000	0.01182	0.00000	0.00000
0011	0.00000	0.00000	0.00705	0.01094	0.03009	0.03099
1110	0.00937	0.00937	0.00168	0.00000	0.31846	0.33264
1101	0.03065	0.03065	0.03117	0.00033	0.03250	0.02790
1011	0.25295	0.25295	0.00576	0.00000	0.12987	0.12314
0111	0.00000	0.00000	0.00000	0.00978	0.01353	0.00982
1111	0.60190	0.60190	0.85770	0.89291	0.31846	0.33264

The results showed that the probabilities for the mastery of all four skills were 60.1% under DINA and DINO models, while it was 85.7% using a mixed model. The results for the validated Q-matrix showed the opposite pattern. The DINA model had the highest probability of 89.2%, while the probabilities for the DINO and mixed models were 31.8% and 33.2%. Finally, Table 5 compares the three DINA, DINO and mixed models across the original and validated Q-matrices. The results revealed that the validated DINA model displayed a significantly better fit (χ^2 = 2.79, p = .000) than the original model. It yielded lower AIC (23176 vs 23179) and BIC (23569 vs 23572) indices.

Table 5

Comparison of Models

Model	Q-Matrix	Log-like	Deviance	AIC	BIC	Chi-sq.	df	p
DINA	Original	-11494.000	22989.000	23179.000	23572.000	2.796	0.000	0.000
	Validated	-11493.000	22986.000	23176.000	23569.000			
DINO	Original	-11494.000	22989.000	23179.000	23572.000	11.900	0.000	0.000
	Validated	-11488.000	22977.000	23167.000	23560.000			
Mixed	Validated	-11488.000	22976.000	23166.000	23559.000	10.860	0.000	0.000
	Original	-11482.000	22965.000	23155.000	23548.000			

The validated DINO model showed a significantly better fit (χ^2 = 11.90, p = .000) than the original model. It enjoyed lower AIC (23167 vs 23179) and BIC (23560 vs 23572) indices. Unlike the previous two models, the original model showed a significantly better fit (χ^2 = 1086, p = .000) than the validated model; it yielded lower AIC (23155 vs 23166) and BIC (23548 vs 23559) indices.

Discussion

The study set out to investigate the cognitive subskills and test specific facets of the IELTS listening test. However, before dealing with the details of the discussion, it is worth mentioning that the findings from CDA modelling and the related results will be most useful when teachers, and tester are able to interpret the results. To start with, the original and validated Q-matrices for a 40–item IELTS listening test were first created to answer the first research question. To examine the fit for the models, the first research question was investigated. Based on the original Q-matrix,

the mixed DINA-DINO model was the best model having a more suitable fit. The validated matrix was the second-best and based on the validated matrix, and the DINO model was the third-best. The DINO model based on the original model was considered to have the least fit. However, regarding the absolute fit index of chi-square (MAX2), the mixed model enjoyed the lowest MAX2 statistic (Table 2); the MADcor, MADq3 and SRMSR indices indicated a good fit. Therefore, strong evidence in the study supported the fit of the models. Fit-wise, the study is consistent with the findings by Yi (2017). The model is good means that it is recommended for analyzing items in various settings in order to check its usefulness in numerous research contexts for various purposes.

Examining the second research question (Table 3), the mixed method showed the highest probability of mastery using the original and validated Q-matrix: The skill probabilities for multiple choice (92.7 %), diagram labelling (92.5 %) and short answer (95.2 %) were the highest probabilities. All three methods showed an acceptable agreement on diagram labelling (87 %). This finding implies that in a compensatory model, mastery of some skills will compensate for the non-mastery of other skills (e.g., Lei & Li, 2016; Rupp & Templin 2008; Wang & Gierl, 2011), while in non-compensatory models, the lack of mastery of one or more skills is not compensated by other skills. Here, the role of feedback in sub-test design and development and in teaching the listening skill can be more noted. The test designers can use the result of the CDM-related research findings to improve the underlying structure of the test. This demands the phenomenon of feedback application and a need to inform the test takers or language learners of the consequences of their performance and what they have and have not mastered, what they need to improve and what they do not. Min and He's (2021) findings support the efficiency of CDM-driven feedback related to the listening skill; they indicate that each test taker's individual ability and cognitive performance and perception of a certain level of oral texts are well-diagnosed through CDM; CDM can provide more individualized feedback for the teachers because they will be aware of the process of the learning and instruction with reference to the performances of the test takers on various item types on the IELTS listening test.

To answer the third research question (Table 4), the skill pattern probabilities show the percentages of mastery of combinations of skills; DINA and DINO models showed that the probabilities for non-mastery of all four skills were 8.94%, and under the mixed model, it was 5.77%. However, based on the validated Q-matrix, the DINA model had the lowest probability of 2.58%, while the DINO and mixed models were 8.57% and 8.48%. Along with this finding, Aryadoust's (2018) finding should be considered; he indicates that listeners' non-mastery of listening elements cannot be compensated for other constituents. The results of the last research question showed that the validated DINA model showed a significantly better fit ($\chi^2 = 2.79$, $p = .000$) than the original model.

These all boil down to the fact that the results of assessment can dramatically affect all the stakeholders involved in the educational context, i.e., the teachers, the testers, the counsellors, and the curriculum designers (Fulcher, 2018). In particular, when it is related to a high-stakes test, its impact can be more global and widespread (Fulcher, 2020, 2021). To be more specific, if we are supposed to have an extensive and acceptable modification in education or training of whatever

kind, we need to obtain feedback from the process of instruction and learning. One of these diagnostic tools can be exams, such as achievement tests, summative tests, or formative assessments and assessment for learning (Stigler, 2010). IELTS seems to play a significant role in bringing about positive change at the macro-level to education and at the micro-level to the individuals, i.e., teachers, students and assessors. Therefore, it is implied that educators need diagnostic information that would facilitate the rapid progress of change to the life chances of IELTS candidates.

Looking back at the findings resulting from the skill probabilities examined using the three models of DINA, DINO, mixed DINA, and DINO can lead the test designers to take some required initiative as to the omission or removal of some items. As a diagnostic tool, this can impact on the teachers and learners as well as test designers. That is to say, when the test designers remove some sections or change the content of a test, it automatically affects the educational elements. The approach adopted in this exploratory study offers a way that could be used to focus discussions on language learners' strengths and weaknesses. The finding is supported by Fulcher (1991, 2020, 2021). He indicates that the feedback taken from the preparation courses can be used as a diagnostic tool for helping the learners to develop their communicative abilities.

As a brief look back at the findings (Tables 2 and 4) indicates, the mixed method is the best to diagnose the language learners' strengths and weaknesses concerning their mastery or non-mastery of the special content. Therefore, since listening skills are both compensatory and non-compensatory, knowledge of some contents of the listening skill can compensate for the lack of mastery of other listening content. Occasionally, knowledge of some listening skills may not compensate for the lack of mastery of other skill(s). This might be of importance for teachers and test developers. Further research is also required to identify the compensatory and non-compensatory aspect of listening skills.

On the other hand, the results of Table 1 and Table 2 have implications for test designers and teachers. When it is statistically approved that the mixed DINA-DINO model using the original Q-matrix was the best model and the items on the test had a good fit, both indicate that testers can rely on this model when they design test items, as this model can provide them with more reliably diagnostic item-relevant feedback. Additionally, it has some implication for the teachers. They can put more confidence in the test as an indicator of listening proficiency. This, in turn, can help the teachers in preparation courses. This is supported by Fulcher (2006, 2013a, 2013b). He indicates that in test design and test retrofitting, test purpose should be taken into account because on the basis of the purpose, test specifications or blueprints and test items can be developed and triggered to the needs of the test takers. In this case, a test can be relatively a valid measure of the intended trait.

No test and no educational materials can be perfect; to support this statement, Table 2 indicates that the IELTS listening test seems to be a relatively acceptable indicator of listening comprehension proficiency as assessed by the University of Cambridge. However, no generalization is made, as various research types and numerous sources of validity with much larger sample size in various EFL and ESL setting are required to deal with the effectiveness, efficiency, practicality and validity argument of IELTS (Alavi et al., 2018). Since IELTS is an

international test, it has an international influence and implications, and hence, an investigation into its deep structure for potential future application is required. More remarkably, this indicates that there seems to be a close line between the IELTS listening construct and the educational demands of the real world. Admittedly, these demands have occasionally left the high-stakes tests open to debate. In this connection, Fulcher (2018) indicates that considering the test purpose, test retrofit and test consequences, it is natural for high-stakes tests, such as IELTS, to be challenged.

Whatever bridges the educational demands and real-world requirements are assessment tools. That is to say, we need modern assessment tools that can contribute to improving English language education and high-stakes tests, such as IELTS. To detect what happens in the process of test development, assessing the items with the use of CDM can diagnose the potential faults; since the diagnosed traits will be fed back to the process of teaching, and learning and this will be effective for learner-oriented assessment, or formative assessment (Fulcher, 2020, 2021)

Internationally, listening passages on the IELTS listening test are played only once so as to resemble real-world listening. It is designed and administered in a read- listen-write fashion (Field, 2005), so that IELTS candidates are committed to noticing the three skills of listening, reading and writing at the same time. Therefore, it is not easy to process the intended information and is also overwhelming in connection with teaching the IELTS listening test in IELTS preparation courses. On the one hand, this needs further use of modern assessment tools to find out the problems with a particular test. On the other hand, it can help the teachers and the testers diagnose construct irrelevant variances (Messick, 1995). That is to say, some variances irrelevant to the nature of the test and instruction might interfere with the performances of the examinees on the real test, which can affect the way they learn. In other words, cognitive variables can exert an irrelevant impression on the IELTS candidates' future chances for academic purposes or social and communicative objectives. These all need diagnostic information to improve the test and inform the IELTS candidates of their strengths and weaknesses.

More significantly, IELTS has also made reference to test preparation courses; however, this educational agenda, i.e., teaching to the test, has been criticized by some scholars (Gipps, 1994). All over the world, there lies a flux of immigration and academic study in English spoken countries, which provides an impetus to the necessity of conducting educational courses named IELTS preparation courses. These IELTS-readiness requirements oblige the language learners and the course instructors to create some roadmap for preparing IELTS candidates to obtain the score they need for whatever purpose they take the test. The listening section of the IELTS is one of the skills which demands further research and investigation. In this connection, Fulcher (2010) indicates that teaching to the test should not focus on practicing test items rather it should develop communicative skills and abilities which will automatically boost the score, too.

Conclusions

Based on the discussion above, the inferences and the implications resulting from IELTS can be multi-faceted. Educators in IELTS preparation courses should be aware of the fact that this construct is more demanding and in the real world, it might cause communicative problems. So, educators are advised to spend more time on teaching this sub-skill. Furthermore, it can affect the

language learners themselves through a lack of awareness of the easiness or difficulty of items or sub-skills. Added to this, the test and curriculum developers and course designers can also take advantage of the findings obtained from CDMs, as these models can help create a comprehension of the mastered and non-mastered content.

The results obtained from running the three models on the test items indicate that the black box of language testing is unknown to the test takers and the IELTS preparation course instructors. However, IELTS is an international assessment tool with international implications for education at the macro level and individuals at the micro-level. With the use of new psychometrical tools, it can be more feasible to motivate test designers to examine the deep structure of the test concerning psychometrically contributive instruments. This has the pedagogical implication that the cycles of education, i.e., teaching, testing and learning, and research, should run in parallel to bring about a productive educational result. That is to say, when the findings of CDM-driven information go into teaching, a context of the interaction is created, and the instructional and learning elements cooperate. In the end, in generalizing the findings of our study, we are cautious, as the study needs to be further researched in other contexts. There is no claim that the findings of the present study are inclusive in terms of the choice of the best assessment tool and the group of students we have investigated.

References

Alavi, S. M., Kaivanpanah, S., & Panahi, A. (2018). Validity of the listening module of International English Language Testing System: multiple sources of evidence. *Language Testing in Asia, 8*(8). https://doi.org/10.1186/s40468-018-0057-4

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

Aryadoust, V. (2018). A Cognitive Diagnostic Assessment Study of the Listening Test of the Singapore–Cambridge General Certificate of Education O-Level: Application of DINA, DINO, G-DINA, HO-DINA, and RRUM, *International Journal of Listening.* https://doi.org/ 10.1080/10904018.2018.1500915

Aryadoust, V. (2019). A review of comprehension subskills: A Scientometrics perspective. *System, 88*(2). https://doi.org/10.1016/j.system.2019.102180

Badger, R., & Yan, X. (2006). The Use of Tactics and Strategies by Chinese Students in the Listening Component of IELTS. *IELTS Research Reports, 9*(2).

Bodie, G. D., & Worthington, D. L. (2010). Revisiting the listening styles profile (LSP-16): A confirmatory factor analytic approach to scale validation and reliability estimation. *The International Journal of Listening, 24*(2), 69-88. https://doi.org/ 10.1080/10904011003744516

Cambridge IELTS 11. (2016). *Cambridge IELTS 11: Official examination papers from of Cambridge: ESOL Examinations.* Cambridge: Cambridge Publications.

Cambridge IELTS 12. (2017). *Cambridge IELTS 12: Official examination papers from University of Cambridge: ESOL Examinations.* Cambridge: Cambridge Publications.

Carless, D., & Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 1315-1325. https://doi.org/10.1080/02602938.2018.1463354

Charge, N., & Taylor, L. (1997). Recent developments in IELTS", *ELT Journal, 51*(4), 374-380. https://doi.org/10.1093/elt/51.4.374

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

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology, 54*, 304-313. https://doi.org/10.1016/j.jmp.2010.01.001

DeCarlo, L. T. (2012). Recognising uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36*, 447-468. https://doi.org/10.1177/0146621612449069

De La Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*, 343-362. https://doi.org/10.1111/j.1745-3984.2008.00069.x

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

De La Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333-353. https://doi.org/10.1007/BF02295640

Field, J. (2005). The Cognitive Validity of the Lecture-based question in the IELTS Listening Paper. *IELTS Research Reports, Volume, 9*, 17-65.

Fulcher, G. (1991). The role of assessment by teachers in schools. In T. Caudery (Ed.), *New Thinking in TEFL (The Dolphin Series, No. 21)*. Denmark, Aarhus University Press, 138-158.

Fulcher, G. (2006). Test architecture. Journal Foreign Language Education Research, 9.

Fulcher, G. (2010). *Glenn Fulcher Talks to ELT News*. Athens, Greece.

Fulcher, G. (2012). Assessment Literacy for the Language Classroom. *Language Assessment Quarterly, 9*(2), 113-132. http://dx.doi.org/10.1080/15434303.2011.642041

Fulcher, G. (2013a). Test Design and Retrofit. In C. A. Chapelle (2012). *The Encyclopedia of Applied Linguistics* (pp. 5809-58-17). Malden MA: Wiley Blackwell https://doi.org/10.1002/9781405198431.wbeal1199

Fulcher, G. (2013b). Language testing in the dock. In A. J. Kunnan (Ed.), The Companion to Language Testing. London: Wiley-Blackwell.

Fulcher, G. (2018). *Glenn Fulcher Talks to Bahram Behin*. JALDA, Iran

Fulcher, G. (2020). Operationalizing Language Assessment Literacy. In D. Tsagari (Ed.), *Language Assessment Literacy: From Theory to Practice* (pp. 8-28) Cambridge Scholars.

Fulcher, G. (2021). Language Assessment Literacy in a Learning-Oriented Assessment Framework. In A. Gebril, (Ed.), *Learning-Oriented Language Assessment: Putting Theory into Practice* (pp. 254-270). New York: Routledge. https://doi.org/10.4324/9781003014102

George, A. C., & Robitzsch, A. (2015). Cognitive Diagnosis Models in R: A didactic. *The Quantitative Methods for Psychology, 11*, 189-205. https://doi.org/10.20982/tqmp.11.3.

George, A. C., Robitzsch, A., Kiefer, T., Grob, J., & Unlu, A. (2016). The R Package CDM for Cognitive Diagnosis Models. *Journal of Statistical Software, 74*(2), 1-25. https://doi.org/ 10.18637/jss.v074.i02

Gipps, C. V. (1994). *Beyond Testing: Towards a Theory of Educational Assessment.* London: Routledge.

Hamid, M. O. (2016). Policies of Global English Tests: Test-Takers' Perspectives on the IELTS Retake Policy. *Discourse: Studies in the Cultural Politics of Education, 37*(3), 472-487. https://doi.org/10.1080/01596306.2015.1061978

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles, *Languag Testing, 32*(3), 317-336. https://doi.org/10.1177/0265532214564505

Hattie, J., & Timperley, H. (2007). "The Power of Feedback." *Review of Educational Research, 77*(1), 81-112. https://doi.org/10.3102/003465430298487

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

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement, 49*, 59-81. https://doi.org/10.1111/j.1745-3984.2011.00160.x

Lei, P. W., & Li, H. (2016). Performance of Fit Indices in Choosing Correct Cognitive Diagnostic Models and Q-Matrices. *Applied Psychological Measurement, 40*(6), 405-417. https://doi.org/10.1177/0146621616647954

Leighton, J. P. (2019). Students' interpretation of formative assessment feedback: Three claims for why we know so little about something so important. *Journal of Educational Measurement, 56*(4), 793-814. http://doi.org/10.1111/jedm.12237

Li, H., & Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment, 18*(1), 1-25. https://doi.org/10.1080/10627197.2013.761522

Li, X., & Wang, W. C. (2015). Assessment of Differential Item Functioning Under Cognitive Diagnosis Models: The Dina Model Example. *Journal of Educational Measurement, 52*(1), 28-54. https://doi.org/10.1111/jedm.12061

Messick, S. (1995). Validity of Psychological Assessment: validation of inferences from" 'person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

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

Nakatsuhara, F., Inoue, C., & Taylor, L. (2017). An investigation into double-marking methods: comparing live, audio and video rating of performance on the IELTS Speaking Test, *IELTS Research Reports, Online Series 1*

Phakiti, A. (2016). Test-takers' performance appraisals, appraisal calibration, state-trait strategy use, and state-trait IELTS listening difficulty in a simulated IELTS Listening test. *IELTS Research Reports Online Series*, *6*, 1-3.

Pilcher, N., & Richards, K. (2017). Challenging the power invested in the International English Language Testing System (IELTS): Why determining 'English' preparedness needs to be undertaken within the subject context. *Power and Education, 9*(1), 3-17. https://doi.org/ 10.1177/1757743817691995

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

Robitzsch, A., Kiefer, T., George, A., C., & Uenlue, A. (2014). CDM: Cognitive Diagnosis Modeling. R Package Version 4.1. http://CRAN.Rproject.org/package=CDM

Rupp, A. A., & Templin, J. (2008). Effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement, 68*(1), 78-96. https://doi.org/10.1177/0013164407301545

Sorrel, M. A., Abad, F. J., Olea, J., De La Torre, J., & Barrada, J. R. (2017). Inferential Item-Fit Evaluation in Cognitive Diagnosis Modeling. *Applied Psychological Measurement, 41*(8), 614-631. https://doi.org/10.1177/0146621617707510

Stigler, J. W. (2010). *Formative Assessment*. Corwin: The USA Templin, J., & 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

Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and Learning Second Language Listening: Meta-cognition in Action.* New York: Routledge.

Wang, C., & Gierl, M. J. (2011). Using the Attribute Hierarchy Method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement, 48*(2), 165-187. https://doi.org/10.1111/j.1745-3984.2011.00142.x

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

Acknowledgements

Not applicable.

Funding

Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.