Alavi et al. Language Testing in Asia (2018) 8:8
https://doi.org/10.1186/s40468-018-0057-4

Language Testing in Asia

RESEARCH Open Access

Validity of the listening module of international English language testing system: multiple sources of evidence

Sayyed Mohammad Alavi*, Shiva Kaivanpanah and Ali Panahi Masjedlou

* Correspondence: smalavi@ut.ac.ir
Faculty of Foreign Languages and Literatures, University of Tehran, Tehran, Islamic Republic of Iran

Abstract

Background: A critical issue in International English Language Testing System (IELTS) lies in the significance of the validity of IELTS listening comprehension test (hereafter IELTS LCM). However IELTS listening validity has been investigated, it has not been investigated with reference to multiple sources of evidence regarding item internal factors. To bridge this gap, we investigated its construct validity with use of structural equation modelling (SEM) and assessed differential item functioning (DIF) through cognitive diagnostic modelling (CDM) and Mantel Haenszel (MH).

Methods: In this study, first, the participants signed a consent form for participation in the study; then, 480 participants were administered a proficiency test designed by the university of Cambridge; next, out of 480 participants, 463 participants were administered a 40-item IELTS LCT developed by the University of Cambridge. Finally, the data were analyzed with use of LISREL for probing the construct validity of the test; also, for detecting the potential DIF items, MH and CDM were used to make the results of DIF related findings more reliable.

Results: The results of the first study confirmed an appropriate model fit, so that all four constructs, i.e., gap filling, diagram labelling, multiple choice and short answer on IELTS LCT, had a statistically significant contribution to IELTS LCT. However, construct-related evidence may not lead to the whole validity. This given, the second study examined the DIF items to argue the validity of IELTS LCT: MH detected 15 DIF items and CDM detected at least 6 DIF items and at most 12 DIF items.

Conclusions: Due to its international nature and world-wide evaluative contribution, IELTS needs to have approximately (not absolutely) a stable factor structure, so that it should be invariant across populations and various cultures. More naturally, a test highly valid in one context might suffer from some degree of invalidity with some related constructs in another context. This in mind, our perspective in this research is not recommended to be taken as a one-size-fits-all model: Neither generalization nor claim is made based on the present study.

Keywords: SEM, Validity, MH, CDM, DIF

Background

International English Language Testing System (IELTS) is an admission requirement for either immigration or education abroad and focuses on language use in a social and academic context (Nakatsuhara et al. 2017; Phakiti 2016). In design terms, IELTS

listening comprehension test (LCT) is intensive, i.e., played just once; it is also in a read-listen-write format (Field 2005). This is overwhelming as the learners are obliged to pay simultaneous attention to three skills: listening, reading and writing, so it is demanding in format for processing information; this under-represents IELTS listening construct (Aryadoust 2012).

To date, more research has been conducted on listening in second or foreign language (Alavi and Janbaz 2014; Bodie and Worthington 2010; Harding et al. 2015; Kimura 2016; KÖk 2017; Roussel et al. 2017; Vandergrift 1997, 2006, 2007). In particular, more and more research has been conducted on IELTS since the IELTS research program started in 1995, so that more than 110 empirical studies have received grant so far (Nakatsuhara et al. 2017); however, to date, there has been a paucity of research in the field of IELTS listening and just fewer bodies of research (e.g., Aryadoust 2011, 2012, 2013; Badger and Yan 2006; Field 2005; Harding et al. 2015; Phakiti 2016; Winke and Lim 2014) have been conducted on IELT LCT as far as we are aware of.

The current study is on the (construct) validity of IELTS LCT; construct validity is a crucial element for language testing or large scale public tests (Cronbach and Meehl 1955; Kane 2013, 2016). Over the course of the years, some distinguished language testing scholars (Kane 2013, 2016; Messick 1974, 1986, 1995, 1996; Newton and Shaw 2015; Newton and Baird 2016. Sireci 2017) have accepted the evidence-based definition of validity as a unitary concept (Messick 1974): it refers to the meaningfulness, usefulness, and appropriateness of the degree evidence and theory weaken or support the inferences and decision made based on a test; however, to judge whether the construct of a test (IELTS LCT, for example) is valid or not requires a validation procedure and a multiple source of evidence. The researchers agree with various types of evidence such as test content, response processes, internal structure, relation to other variables, and test consequence (Sireci 2017). To this end, this study first examined the construct validity of IELTS LCT with the use of structural equation modeling (SEM), i.e., confirmatory factor analysis (CFA) with the use of LISREL software; then, in phase 2, it assessed differential item functioning (DIF) with the use of cognitive diagnostic modeling (CDM) and Mantel-Haenszel (MH) method; the reason why we used two methods for item analysis was to put more confidence in the accuracy of DIF findings.

Literature review

SEM and factor analysis

Factor analysis refers to a multivariate technique (Sawaki 2012; Schmitt 2011) required for making an interpretation of a large number of correlations (Field 2009; Khine 2013); it is a statistical method used for testing and estimating the relations (Alavi and Ghaemi 2011; Ockey and Choi 2015) inherent in a group of variables in order to gain insight into the underlying causal processes (In'nami and Koizumi 2011; Kunnan 1994, 1998).

A look at literature review reveals more studies conducted with use of SEM (Alavi and Ghaemi 2011; Alavi et al. 2011b; Cai 2013; Carr 2006; Phakiti 2008, Sawaki et al. 2009; Schoonen 2005; Song 2008). However, very few bodies of research have been conducted on IELTS LCT with the use of SEM. A very recent SEM study associated with IELTS listening has been done by Phakiti (2016); his findings suggested that there are complex structural relationships among test-takers'confidence, calibration, trait,

strategy use, IELTS listening difficulty, and performance on IELTS listening. Another study on IELTS listening was done by Field (2005); his study explores the cognitive validity of lecture-based questions in IELTS LCT; his findings support the cognitive validity of the IELTS. In the same vein, Badger and Yan (2006) did a research on IELTS listening strategies and their findings supported the construct validity of IELTS listening, too.

Differential item functioning

DIF exists when different groups of learners have different probability of successfully answering an item (Drabinova and Martinkova 2017; Ferne and Rupp 2007; Li and Wang 2015); therefore, if the test takers have less or more the same knowledge, then they should perform similarly on test items; DIF is needed for test validity and test fairness (Fidalgo et al. 2014; Hou et al. 2014; Pae 2004, 2012; Su and Wang 2005; Zumbo 2003, 2007).

A look at literature review indicates an abundant number of DIF studies; the studies appear in various DIF-related factors, such as gender (e.g., Abbott 2006; Amirian et al. 2014; Aryadoust 2012; Li and Suen 2013; Pae 2012; Rezaee and Shabani 2010; Song et al. 2015), age (e.g., Geranpayeh and Kunnan 2007), academic background (e.g., Alavi et al. 2011a; Pae 2004), text familiarity (e.g., Ahmadi and Jalili 2014), field of study (Barati et al. 2006), and language background (e.g., Harding 2011; Kim 2001; Kim and Jang 2009). As noted in introduction, among all these studies, a study exactly related to the DIF of IELTS LCT was conducted by Aryadoust (2012), as far as we are aware; his research indicates some construct-underrepresentation on IELTS LCT. Therefore, this study is in line with DIF detection.

Two DIF-detection methods: MH and CDM

MH statistic (Mantel and Haenszel 1959) is one of the most globally utilized procedures for DIF detection, as it is relatively easy to calculate; it does not need large sample sizes; it includes a test of statistical significance and also reports effect size (Monahan and Ankenmann 2005, Monahan and Ankenmann 2010; Su and Wang 2005). That said, the Mantel-Haenszel statistic makes comparisons of item performance for various groups; it compares examinees of similar proficiency levels, instead of comparing overall group performance on an item (Michaelides 2008). That said, it needs however to be acknowledged that MH does not behave optimally in all situations and that this might lead to an error in DIF detection (Guilera et al. 2013).

Another method is CDM; it is a psychometric model developed for assessing examinees mastery and non-mastery of skills or attributes (Chen et al. 2013; de la Torre 2011); recently, various kinds of CDMs are used, such as deterministic inputs, noisy and gate model (DINA; Junker and Sijtsma 2001) and the deterministic inputs, noisy or gate model (DINO; Templin and Henson 2006). As for the significance of CDM, George and Robitzsch (2014) recommend the use of CDM as one of the recent statistical tools for detecting DIF and plenty of psychometric questions in relation to DIF can be addressed with use of CDM (Hou et al. 2014). To date, only a few studies have been conducted on DIF assessment within the framework of CDM (Drabinova and Martinkova 2017; Li and Wang 2015; Hou et al. 2014; Li 2008; Zhang 2006). However an extensive body of research has been done in the area of cognitive diagnosis of students' learning (Li and Wang 2015; de la Torre 2011; de la Torre and Douglas

2004; Junker and Sijtsma 2001), no study has so far been done on detecting the DIF of IELTS LCT with use of CDMs, so that some researchers (e.g., George and Robitzsch 2014) suggest the use of CDM for DIF detection.

Research questions

Based on the review of literature, this study investigates the following research questions:

RQ1: Does the factor structure of IELTS LCT reflect the design of the test in terms of task types, i.e., gap filling, diagram labeling, multiple choice, and short answer?

RQ2: Does group membership (gender) exert any bias towards the participants' performance on the items of IELTS LCT as investigated by Mantel-Haenszel (MH)?

RQ3: Does group membership (gender) exert any bias towards the participants' performance on the items of IELTS LCT as investigated by Cognitive Diagnostic Modeling (CDM)?

Method

Participants and context

The study was carried out at various English Language Institutes in Iran; these institutes mainly aim at administering monthly IELTS mock-tests to the potential IELTS candidates. Also, the participants were mostly on IELTS preparation courses; the participants in both phases of the present study were those who needed to attend IELTS preparation course. As for sample size, it determines the quality of SEM study (Ockey and Choi 2015); the minimum and maximum sample size for SEM is indicated to be 100 to 150 subjects (Ding et al. 1995; Khine 2013) and 400 subjects (Boomsma 1987), respectively, or 5–10 subjects for every item or variable (Bentler and Chou 1987).

Therefore, in this study, 480 participants took a proficiency test adopted from Cambridge IELTS books; the performances of 17 participants were excluded from this study, as their performances were of extraneous variances. Finally, an adequate number of 463 participants (Table 1) took part in the study; they had studied the English language (for an ultimate goal of passing IELTS) for approximately 4 years; they were characterized by the same cultural, societal, native language, and educational context. The researchers strove to obtain access to real data of IELTS LCT; however, due to confidentiality reasons associated with IELTS organization, it was not possible. As such, the participants took the test in IELTS mock-test condition. Also, 18 teachers (8 female teachers and 10 male teachers) all majoring in ELT and teaching IELTS preparation course took part in the study; five of them had a B.A. in English language, 9 of them had an M.A. in TEFL, and four of them were PhD candidates in TEFL.

Table 1 Demographic data of participants

Number	Groups			Items	
	Female	Male	Age	Items	Type of item
463	227 (49.03%)	236 (50.97%)	19–25	40	Dichotomous

Materials

Two IELTS LCTs adopted from IELTS test books (Cambridge IELTS 2016; 2017) were used: a proficiency test and a main test; the first was used for proficiency purpose and the second was used to probe the (construct) validity of IELTS LCT. IELTS LCTs were played in 30 min, and the participants were given 10 min to transfer their answer to the answer sheet (IELTS Handbook 2007). Also, a summarized handout was used for strategy instruction adapted from Tips for IELTS (McCarter 2006), Action Plan for IELTS (Jakeman and McDowell 2006), and Step Up to IELTS (Jakeman and McDowell 2004). These techniques and strategies were instructed in five sessions in context and with related listening subsections, related to IELTS LCT in five sessions. Since IELTS LCT demands its own strategies (McCarter 2006; London Teacher Training College 2005) and testwiseness also maximizes the performance on test (Rogers and Yang 1996), so the questions on real IELTS tests are susceptible to testwiseness strategies. This would approximately keep the test takers' condition on mock-test setting similar to the real test takers' situation on real test. The specifications of the main test appear below (Table 2).

Procedures and data analysis

First, the participants signed a consent form for participation in the study. Then, a proficiency test, i.e., IELTS LCT, was administered and the reliability of the measurement tool was investigated; that is to say, we ran Cronbach's Alpha on IELTS LCT which reached at a reliability of 0.66 (0.73 and 0.58 for the males and females, respectively). Of course, on all of the 13 volumes of Cambridge IELTS Books appears a phrase which reads *authentic papers*, which was the main impetus for this investigation.

Next, the teachers instructed the strategies and finally, the main test was administered. We ran confirmatory factor analysis using LISREL software to probe the construct validity of the test. We also analyzed the data for DIF detection related to gender using MH and CDM.

Results and discussion

Phase 1

Data analysis showed that the performances of the participants on proficiency test and on the main test (M = 22.94, SD = 4.62; M = 23.34, SD = 5.06), respectively, were approximately the same (Tables 3 and 4). As it is clear from Table 3, 480 participants took a proficiency test, and just 463 participants' performances (Table 4) on items of the main test were analyzed for the purpose of confirmatory factor analysis and item bias.

If the absolute values of the skewness and kurtosis statistics are lower than 2, the univariate normality of the items is met (Bae and Bachman 2010). As it is evident

Table 2 Specifications of the main test

Social needs on IELTS listening		Educational needs on IELTS listening	
GF	DL	MC	SA
14 items	6 items	10 items	10 Items
1–14	15–20	21–30	31–40

GF gap filling, *DL* diagram labeling, *MC* multiple choice, *SA* short answer

Table 3 Descriptive statistics for the proficiency test

Number	Minimum	Maximum	Mean	SD
480	13	38	22.94	4.62

in Table 5, this assumption was met. Also, the Mardia test of multivariate normality of − 6.31 was lower than 1680, so the assumption of multivariate normality was also met. This was calculated with this formula: $p \times (p + 2)$ or $40 \times (40 + 2) = 1680$ (Khine 2013); here, p stands for the number of observed variable which was 40 in this study.

Figure 1 displays the 40 items (the items in squares) of IELTS LCT. Four sub-sets of items, i.e., Gap filling (GF), diagram labelling (DL), multiple choice (MC), and short answer (SA), measure four latent variables (the four ovals), which eventually, measure total IELTS LCT (the oval titled listen). Based on the statistical analysis outlined in Fig. 1 and Table 6, among the 14 items of the GF, eight (items 4 to10 and 12) were significant, i.e., = > .30 (higher than .30). Five of the six items on DL (items 16 to 20) were higher than .30. And also, eight items (items 21, 23, and 25 to 30) of MC were significant. Finally, just three items (items 31, 32, and 37) of SA were significant. The four latent variables of gap filling (b = 1.10), diagram labeling (b = .43), multiple choice (b = .60), and short answer (b = .91) all had significant contributions to the total IELTS LCT (Fig. 1).

As seen in Tables 6 and 7, the results of the chi-square (χ^2 (736) = 1226.49, p = .000) indicated the poor fit of the model. However, chi-square is sensitive to sample size (Hooper et al. 2008) that is why its ratio over the degree of freedom (1226.49/736 = 1.66) should be consulted. Since this ratio is lower than 3, it can be concluded that the overall model enjoys a good fit.

As Table 7 reveals, the root mean square of error approximation (RMSEA) of .038 and its 90% confidence intervals, i.e., [90% CI (.034, .042)] which were lower than .05 as well as the closeness of fit statistic (PCLOSE) which was higher than .50 supported the fit of the model. Further evidence which confirmed the fit of the model results from the non-normed fit index (NNFI = .91), comparative fit index (CFI = .92), incremental fit index (ICI = .92), and goodness of fit index (GFI = .90), all of which were equal to or higher than .90. Also, the critical N (CN = 312.97) which was higher than 200, indicating the sampling adequacy of the model supported the fit of the model.

Table 4 Descriptive statistics for performance on the main test

Gender		Number	Mean	Std. deviation	Variance
Female	GF	227	8.60	2.21	4.89
	DL	227	4.23	1.48	2.20
	MC	227	6.74	1.95	3.80
	SA	227	4.06	1.79	3.23
	Total	227	23.63	4.54	20.62
Male	GF	236	9.18	2.53	6.41
	DL	236	2.67	1.06	1.14
	MC	236	6.67	1.95	3.83
	SA	236	4.53	1.76	3.10
	Total	236	23.05	5.58	31.17
	Both	463	23.34	5.06	25.89

Table 5 Tests of univariate and multivariate normality

Variable	Min	Max	Skewness	kurtosis
q1	0	1	− 1.54	0.38
q2	0	1	− 0.27	− 1.92
q3	0	1	− 0.49	− 1.75
q4	0	1	− 0.84	− 1.27
q5	0	1	− 0.64	− 1.57
q6	0	1	− 0.91	− 1.15
q7	0	1	0.06	− 1.99
q8	0	1	0.01	− 2.00
q9	0	1	0.10	− 1.98
q10	0	1	− 0.73	− 1.46
q11	0	1	− 1.58	0.49
q12	0	1	− 0.62	− 1.60
q13	0	1	− 1.27	− 0.37
q14	0	1	0.03	− 1.99
q15	0	1	− 0.78	− 1.38
q16	0	1	− 0.42	− 1.81
q17	0	1	0.09	− 1.99
q18	0	1	− 0.09	− 1.99
q19	0	1	− 0.09	− 1.99
q20	0	1	− 0.50	− 1.74
q21	0	1	− 0.37	− 1.86
q22	0	1	− 0.98	− 1.02
q23	0	1	− 0.90	− 1.18
q24	0	1	− 0.45	− 1.79
q25	0	1	− 0.711	− 1.49
q26	0	1	− 1.00	− 1.00
q27	0	1	− 0.53	− 1.71
q28	0	1	− 0.48	− 1.76
q29	0	1	− 1.17	− 0.61
q30	0	1	− 0.77	− 1.40
q31	0	1	0.40	− 1.83
q32	0	1	− 0.01	− 2.00
q33	0	1	0.89	− 1.20
q34	0	1	− 0.60	− 1.64
q35	0	1	0.16	− 1.97
q36	0	1	0.01	− 2.00
q37	0	1	0.64	− 1.57
q38	0	1	0.08	− 1.99
q39	0	1	0.39	− 1.84
q40	0	1	1.01	− 0.97
Multivariate				− 6.31

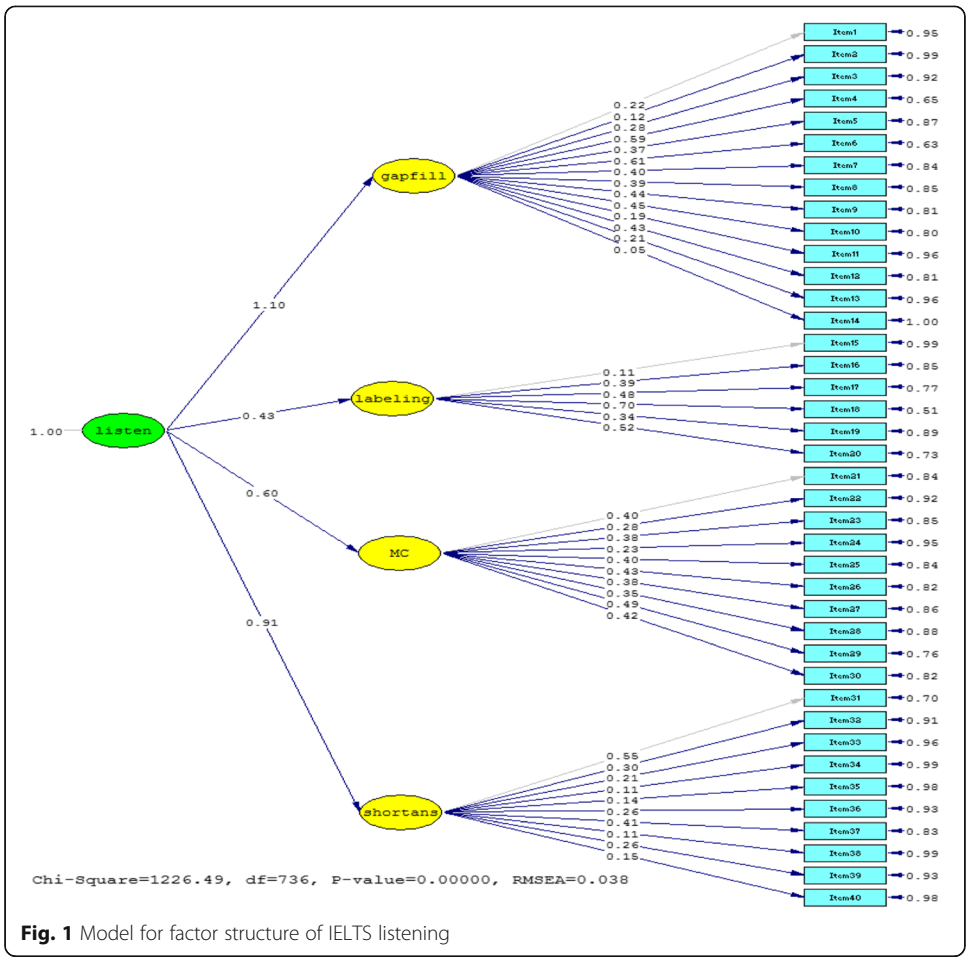

Fig. 1 Model for factor structure of IELTS listening

Table 6 Standardized regression weights of IELTS listening comprehension test

GF		DL		MC		SA	
Items	*b*	Items	*b*	Items	*b*	Items	*b*
1	0.22	15	0.11	21	0.40	31	0.55
2	0.12	16	0.90	22	0.28	32	0.30
3	0.28	17	0.48	23	0.38	33	0.21
4	0.59	18	0.70	24	0.23	34	0.11
5	0.37	19	0.34	25	0.40	35	0.14
6	0.61	20	0.52	26	0.43	36	0.26
7	0.40			27	0.38	37	0.41
8	0.39			28	0.35	38	0.11
9	0.44			29	0.49	39	0.26
10	0.45			30	0.42	40	0.15
11	0.19						
12	0.43						
13	0.21						
14	0.05						

Table 7 Model fit indices

Indices	Model	*p*	Recommended level
Chi-square	1226.49 (736)	.000	Non-significant
Chi-square ratio	1.66	–	= < 3
NNFI	.91	–	= > .90
CFI	.92	–	= > .90
ICI	.92	–	= > .90
GFI	.90	–	= > .90
RMSEA	.038	–	= < .05
95% CI RMSEA	[.034, .042]	–	= < .05
p-close	1.000	–	> .50
CN	312.97	–	= > 200

Phase 2

Differential item functioning with use of Mantel-Haenszel method

Mantel-Haenszel's method (Table 8) showed 15 significant DIF items, of which eight enjoyed large effect sizes, i.e., items 4, 8, 10, and 16 to 20. The effect size values for the Mantel-Haenszel's DIF are weak = 0, moderate = 1, and large = 1.5 (Mantel and Haenszel 1959).

Differential item functioning with use of CDM

As seen in Table 9, the results of the chi-square tests identified 12 DIF items, i.e., items (2, 4, 5, 8 to 10, 14, 18, 20, 21, 26, 33, and 36). One of the main features of CDM DIF is that the original *p* values (fourth column) are recalculated using Holm's adjusted formula, in which the *p* values are penalized for multiple pair-wise comparisons. The Holm *p* value is computed using this formula (Wright 1992, p. 1008): $p\text{-Holm} = 1 - (1 - p)^{K.}$

In this formula, *K* stands for the number of items minus the number of comparison above it; the *K* for the first item is 40, 39 for the second item, 38 for the third, and finally, 1 for the last items. Based on Holm's adjusted *p* values, six DIF items (2, 8, 10, 14, 18, and 20) were flagged.

The unassigned area (UA) is the effect size for the CDM DIF with three values (weak = lower than .059; moderate = .059 to .088; large = higher than .088) (George and Robitzsch 2014, p.414). The results identified 13 items enjoying large effect sizes (items 2, 4, 5, 8, 10, 11, 14, 18, 20, 21, 26, 30, and 36). The summary of the findings appears in Table 10.

Discussion

The purpose of the present study was to investigate the validity of IELTS LCT. In this context of IELTS LCT validation, the hypothetical variables were associated with a construct or test method (gap filling, multiple choice, diagram labelling, and short answer); here, the researchers first hypothesized a model and then examined whether the model is advocated by the present sample. The overall model fit in the phase 1 of the study provided evidence of construct validity for IELTS LCT, so the hypothesized SEM model enjoyed a good fit. What is hence clearly outlined in the analysis is that the individual items revealed to be valid indicators of their assumed factors or constructs, i.e., gap filling, diagram labelling, multiple choice, and short answer.

Table 8 Differential item functioning through Mantel-Haenszel's method

Items	Chi-square	*P*	DIF/non-DIF	Delta MH	Effect size
1	0.72	0.39	Non-DIF	− 0.57	A
2	8.31	0.00	DIF	− 1.41	B
3	2.31	0.12	Non-DIF	0.79	A
4	15.21	0.00	DIF	2.23	C
5	2.38	0.12	Non-DIF	0.83	A
6	4.99	0.02	DIF	1.29	B
7	6.93	0.00	DIF	1.32	B
8	24.67	0.00	DIF	2.48	C
9	1.52	0.21	Non-DIF	− 0.61	A
10	23.81	0.00	DIF	2.60	C
11	5.20	0.02	DIF	1.47	B
12	0.01	0.90	Non-DIF	0.11	A
13	0.81	0.36	Non-DIF	− 0.55	A
14	0.00	0.94	Non-DIF	0.07	A
15	7.39	0.00	DIF	− 1.40	B
16	38.80	0.00	DIF	− 3.04	C
17	55.07	0.00	DIF	− 3.66	C
18	45.80	0.00	DIF	− 3.35	C
19	39.59	0.00	DIF	− 2.84	C
20	16.66	0.00	DIF	− 2.01	C
21	2.62	0.10	Non-DIF	− 0.83	A
22	0.16	0.68	Non-DIF	0.27	A
23	0.06	0.80	Non-DIF	0.18	A
24	0.94	0.33	Non-DIF	0.50	A
25	0.10	0.74	Non-DIF	0.20	A
26	3.48	0.06	Non-DIF	− 1.04	B
27	0.00	0.96	Non-DIF	0.07	A
28	2.23	0.13	Non-DIF	0.76	A
29	1.22	0.26	Non-DIF	0.66	A
30	3.43	0.06	Non-DIF	1.01	B
31	6.25	0.01	DIF	1.27	B
32	0.03	0.84	Non-DIF	0.13	A
33	3.21	0.07	Non-DIF	0.92	A
34	2.12	0.14	Non-DIF	− 0.74	A
35	0.24	0.62	Non-DIF	0.26	A
36	7.60	0.00	DIF	1.33	B
37	3.07	0.07	Non-DIF	0.90	A
38	2.89	0.08	Non-DIF	0.84	A
39	3.65	0.05	Non-DIF	0.95	A
40	0.86	0.35	Non-DIF	0.49	A

The findings of the first phase of the study are consistent with the findings by Phakiti (2016), Badger and Yan (2006) and Zhang (2015) whose findings provide some positive evidence in support of construct validity of the IELTS LCT; the statistical significance

Table 9 Differential item functioning through CDM

Items	Chi-square	Df	*P*	DIF/non-DIF CDM	*P*-Holm	DIF/non-DIF Holm	UA	Effect size
1	2.56	2	0.27	Non-DIF	1.00	Non-DIF	0.07	Moderate
2	17.98	2	0.00	DIF	0.00	DIF	0.14	Large
3	2.53	2	0.28	Non-DIF	1.00	Non-DIF	0.08	Moderate
4	9.03	2	0.01	DIF	0.33	Non-DIF	0.12	Large
5	5.67	2	0.05	Non-DIF	1.00	Non-DIF	0.09	Large
6	5.20	2	0.07	Non-DIF	1.00	Non-DIF	0.07	Moderate
7	5.16	2	0.07	Non-DIF	1.00	Non-DIF	0.08	Moderate
8	22.64	2	0.00	DIF	0.00	DIF	0.17	Large
9	9.84	2	0.00	DIF	0.24	Non-DIF	0.06	Moderate
10	21.78	2	0.00	DIF	0.00	DIF	0.17	Large
11	5.50	2	0.06	Non-DIF	1.00	Non-DIF	0.10	Large
12	1.55	2	0.46	Non-DIF	1.00	Non-DIF	0.05	Weak
13	2.50	2	0.28	Non-DIF	1.00	Non-DIF	0.08	Moderate
14	37.75	2	0.00	DIF	0.00	DIF	0.11	Large
15	0.15	2	0.92	Non-DIF	1.00	Non-DIF	0.06	Moderate
16	0.36	2	0.83	Non-DIF	1.00	Non-DIF	0.01	Weak
17	4.85	2	0.08	Non-DIF	1.00	Non-DIF	0.07	Moderate
18	18.64	2	0.00	DIF	0.00	DIF	0.10	Large
19	1.27	2	0.52	Non-DIF	1.00	Non-DIF	0.05	Weak
20	16.21	2	0.00	DIF	0.01	DIF	0.14	Large
21	7.00	2	0.03	DIF	0.87	Non-DIF	0.10	Large
22	2.26	2	0.32	Non-DIF	1.00	Non-DIF	0.06	Moderate
23	0.09	2	0.95	Non-DIF	1.00	Non-DIF	0.01	Weak
24	2.01	2	0.36	Non-DIF	1.00	Non-DIF	0.06	Moderate
25	1.34	2	0.51	Non-DIF	1.00	Non-DIF	0.01	Weak
26	9.44	2	0.00	DIF	0.28	Non-DIF	0.09	Large
27	1.68	2	0.43	Non-DIF	1.00	Non-DIF	0.01	Weak
28	3.22	2	0.19	Non-DIF	1.00	Non-DIF	0.06	Moderate
29	0.68	2	0.70	Non-DIF	1.00	Non-DIF	0.03	Weak
30	5.78	2	0.05	Non-DIF	1.00	Non-DIF	0.10	Large
31	5.96	2	0.05	Non-DIF	1.00	Non-DIF	0.08	Moderate
32	0.06	2	0.96	Non-DIF	1.00	Non-DIF	0.00	Weak
33	7.44	2	0.02	DIF	0.72	Non-DIF	0.07	Moderate
34	3.32	2	0.18	Non-DIF	1.00	Non-DIF	0.06	Moderate
35	0.06	2	0.96	Non-DIF	1.00	Non-DIF	0.00	Weak
36	11.13	2	0.00	DIF	0.12	Non-DIF	0.1322	Large
37	2.92	2	0.23	Non-DIF	1.00	Non-DIF	0.0547	Weak
38	4.62	2	0.09	Non-DIF	1.00	Non-DIF	0.0792	Moderate
39	4.44	2	0.10	Non-DIF	1.00	Non-DIF	0.0724	Moderate
40	0.40	2	0.81	Non-DIF	1.00	Non-DIF	0.0176	Weak

of construct validity for IELTS LCT is in keeping with the statement that there can be no validity without construct validity (Messick 1974, 1986). However, to argue the validity of a test, we need rich pieces of evidence (Kane 2016; Messick 1974, 1986,

Table 10 Comparison of DIF detection methods

DIF methods	DIF	DIF *p*-adj.	Effect size	
			Moderate	Large
MH	16		10	8
CDM	12	6	16	13

1995, 1996; Sireci 2017), for example, differential item functioning, consistency of the measurements, response processes, internal structure, content, context, test consequence, and cognitive data. Therefore, that our study confirmed the construct validity of the test does not mean that the test is fully valid, as no test is inherently valid or invalid (Sireci 2017); rarely will it be possible for a test to make a prediction of a definite construct (Cronbach and Meehl 1955; Messick 1986); in other words, construct-related evidence may not be the whole validity (Messick 1974, 1986), so no one single piece of evidence for probing the construct validity is sufficient on its own. Clearly, due to the challenging nature of validity, IELTS LCT as a global test with a macro and micro impact needs being viewed and investigated in light of multiple evidences. These given, investigating the degree of validity of IELTS LCT with reference to DIF was also required. That is why phase 2 of the study provided another piece of evidence.

Based on some evidence, IELTS LCT suffers from some degree of invalidity (Aryadoust 2012). Along the same line, in our study, the two methods, i.e., Mantel Haenszel method detected 15 DIF items and CDM flagged at most 12 and at least 6 DIF items (Tables 9 and 10). A closer look at sub-sections reveals all items (six DIF items) of diagram labeling flagged by MH and just two DIF items of diagram labeling were detected by CDM; of course, the difference in the number of DIF items detected by these two methods needs some reflection. That is to say, based on MH, diagram labeling revealed six items (all items) and CDM just two DIF items on diagram labeling. Also, on gap filling, seven DIF items (half of the items) and five DIF items were detected by MH and CDM, respectively.

The findings of study 2 is consistent with Aryadoust's (2012) findings; his research revealed that the first construct in the test was found to be under-represented, as construct under-representation is apparent in the gap filling and diagram labelling in our study too. On the other hand, gap filling (with 14 items) and diagram labelling (with 6 items) both are sub-tests of social dimension of IELTS LCT; however, the relative number of items have not been equally designed (Table 2), as the former has more items than the latter. These all given, it seems that some unwanted or construct irrelevant variances can possibly interfere with sections 1 and 2 of IELTS LCT; they, therefore, need further investigation. As for the analysis of item bias, item-internal evidence for probing the validity is not sufficient on its own too. Therefore, as Cronbach and Meehl (1955) stated, the stability of test scores, i.e., measurement consistency, can be related to construct validation and together with other cognitive and contextual evidence can help with any decision about examining the validity of IELTS LCT.

DIF items can threaten the validity of IELTS LCT; there is some effect-size-based evidence that DIF is not equivalent to bias, but DIF is unavoidable in international tests (Le 2006), such as IELTS; so, not all cases of DIF necessarily have to be interpreted as item bias (Tatsuoka et al. 1988), as the effect size of the DIF item

should be consulted for final decision for either improvement, revision, or removal. Based on the findings from DIF analysis (Tables 9 and 10), we do not claim that the DIF items detected in phase 2 of the study severely pollute IELTS LCT because more study needs for big claims; neither do we suggest the generalization of the findings beyond, as it was done in an Iranian EFL context, where the language learners have the least amount of (or no) exposure to listening input in a social context and in a governmental school setting; they just learn English language at private institutes; also, the learners receive very restricted amount of live audio and visual input from mass media due to some educational policy and governmental decisions in Iranian EFL setting.

As for the construct validity of IELTS, it is played just once (Field 2005); the candidates must pay simultaneous attention to three skills: listening, reading, and writing, as it is in a read-listen-write design. As Aryadoust (2012) maintains, if test takers make use of other skills such as reading or writing beyond the intended skill, (this might pollute the score use and interpretation related to IELTS listening: The bracket is ours). Likewise, some (or most) of the real-world characteristics are missing on IELTS LCT. The listeners perceive the message through scaffolding elements such as lip-reading, facial expression, body language, gestures, and postures. These can underrepresent IELTS listening construct (Aryadoust 2012). This creates another reservation and motivation in line with further investigation into IELTS LCT.

Overall, IELTS LCT seems to be a good indicator of listening proficiency as assessed by the University of Cambridge. To be impressionistic, as the report of the 15-year IELTS teaching experience of the third author of this paper can provoke some thought. With reference to the performance results of hundreds of IELTS candidates undertaking IELTS preparation courses, the IELTS candidates who get a band score of 6 or 6.5 are capable to easily communicate and meet their academic and social needs. This indicates that there seems to be a close line between IELTS listening construct and the demand of real world. Therefore, IELTS seems to be an effective assessment tool; since it sounds to be of global impact, nothing should be taken for granted and more research should be done into it. Of course, as mentioned, the nature of our findings or other researchers' findings and the third authors' impressionistic and personal judgment all need to be more investigated and highly documented. However, the findings of our study call into question Pilcher and Richards' (2017) tone of speech regarding the power of IELTS; their strong claim is that the power of IELTS needs to be challenged; contrary to their findings, our findings indicate that IELTS needs to be more investigated; its invalid sub-parts and sub-constructs need to be improved and revised—and if needed, be removed or replaced—rather than challenged.

Conclusion

In terms of implications, the findings of the study can be thought-provocative; it can motivate the researchers, the materials developers, and IELTS listening test designers and the curriculum designers to be more aware of the psychologically underlying construct of the test. Since, IELTS LCT is an example of a public test that is used to make crucial decisions about huge numbers of people all over the globe, the wash-back effect and the consequential validity of IELTS LCT must be taken into account.

In conclusion, due to its international nature and world-wide evaluative contribution, IELTS needs a stable factor structure, so that it should be invariant across populations and various cultures. More naturally, a test highly valid in one context might suffer from some degree of invalidity with some related constructs in another context. This in mind, our perspective in this research is not recommended to be taken as a one-size-fits-all model and neither generalization nor claim is made based on the present study. The study is limited in scope as the test takers were not real IELTS test takers; they were not also drawn from very large international population. Further research should concentrate on a larger sample size in world-wise educational and cultural contexts, as there is a need to other evidence to warrant further examination of IELTS LCT validity.

Abbreviations

CDM: Cognitive diagnostic modeling; CFI: Comparative fit index; DIF: Differential item functioning; DL: Diagram labelling; GF: Gap filling; GFI: Goodness of fit index; ICI: Incremental fit index; IELTS: International English Language Testing System; LCT: Listening comprehension test; MC: Multiple choice; MH: Mantel-Haenszel; NNFI: Non-normed fit index; RMSEA: The root mean square of error approximation; SA: Short answer; SEM: Structural equation modeling

Acknowledgements

The teachers' help with collecting the data is appreciated. We are thankful to the learners and teachers who participated in this study.

Funding

There is no funding.

Availability of data and materials

Data and materials will be available upon request.

Authors' contributions

All authors made a contribution to this manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 23 January 2018 Accepted: 20 February 2018
Published online: 20 June 2018

References

Abbott, ML. (2006). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing*, *24*(1), 7–36. https://doi.org/10.1177/0265532207071510 .

Ahmadi, A., & Jalili, T. (2014). A confirmatory study of differential item functioning on EFL reading comprehension. *Applied Research on English Language*, *3*(6), 55–68. https://doi.org/10.22108/are.2014.15489 .

Alavi, S. M., & Ghaemi, H. (2011). Application of structural equation modeling in EFL testing: a report of two Iranian studies. *Language Testing in Asia*, *1*(3), 22–35. https://doi.org/10.1186/2229-0443-1-3-22 .

Alavi, S. M., & Janbaz, F. (2014). Comparing two pre-listening supports with Iranian EFL learners: opportunity or obstacle. *RELC Journal*, *45*(3), 253–267. https://doi.org/10.1177/0033688214546963 .

Alavi, SM, Rezaae, AA, Amirian, SMR. (2011a). Academic discipline DIF in an English language proficiency test. *Journal of English Language Teaching and Learning*, *7*, 39–65.

Alavi, SM, Kaivanpanah, S, Nayernia, A. (2011b). The factor structure of a written English proficiency test: a structural equation modeling. *Iranian Journal of Applied Language Studies*, *3*(2). https://doi.org/10.22111/ijals.2011.1008.

Amirian, SMR, Alavi, SM, Fidalgo, AM. (2014). Detecting gender DIF with an English proficiency test in EFL context. *Iranian Journal of Language Testing*, *4*(2), 187–203.

Aryadoust, V. (2011). Validity arguments of the speaking and listening modules of international English language testing system: a synthesis of existing research. *The Asian ESP Journal*, *7*(1), 28–54.

Aryadoust, A. (2012). Differential item functioning in while-listening performance tests: the case of international English language testing system (IELTS) listening module. *International Journal of Listening*, *26*(1), 40–60. https://doi.org/10.1080/10904018.2012.639649.

Aryadoust, V. (2013). Building a validity argument for a listening test of academic proficiency. (PP 1–30). Cambridge: Cambridge Scholars.

Badger, R, & Yan, X. (2006). The use of tactics and strategies by Chinese students in the listening component of IELTS. *IELTS Research Reports*, 9 Retrieved on 15 Sept 2015, from http://www.ielts.org.

Bae, J, & Bachman, L. (2010). An investigation of four writing traits and two tasks across *two languages*. *Language Testing*, *27*(2), 213–234. https://doi.org/10.1177/0265532209349470.
Barati, H, Ketabi, S, Ahmadi, A. (2006). Differential item functioning in high-stakes tests: the effect of field of study. *IJAL*, *19*(2), 27–42.
Bentler, PM, & Chou, CP. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, *16*(1), 78–117. https://doi.org/10.1177/0049124187016001004.
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.
Boomsma, A (1987). The robustness of maximum likelihood estimation in structural equation models. In P Cuttance, R Ecob (Eds.), *Structural modelling by examples*, (pp. 160–188). Cambridge: Cambridge University Press The International Journal of Listening, 24(2), 69–88. https://doi.org/10.1080/10904011003744516.
Cai, H. (2013). Partial dictation as a measure of EFL listening proficiency: evidence from confirmatory factor analysis. *Language Testing*, *30*(2), 177–199. https://doi.org/10.1177/0265532212456833.
Cambridge IELTS (2016). *Cambridge IELTS 11: Official examination papers from University of Cambridge: ESOL examinations*. Cambridge: Cambridge Publications.
Cambridge IELTS (2017). *Cambridge IELTS 12: Official examination papers from University of Cambridge: ESOL examinations*. Cambridge: Cambridge Publications.
Carr, NT. (2006). The factor structure of test task characteristics and examinee performance. *Language Testing*, *23*(3), 269–289. https://doi.org/10.1191/0265532206lt328oa.
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.
Cronbach, LJ, & Meehl, PE. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957.
Ding, L, Velicer, WF, Harlow, LL. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modelling: Multidisciplinary Journal*, *2*(2), 119–143. https://doi.org/10.1080/10705519509540000.
Drabinova, A, & Martinkova, P. (2017). Detection of differential item functioning with non- linear regression: a non-IRT approach accounting for guessing. *Journal of Educational Measurement*, *54*(4), 498–517. https://doi.org/10.1111/jedm.12158.
Ferne, T, & Rupp, AA. (2007). A synthesis of 15 years of research on DIF in language testing: methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, *4*(2), 113–148. https://doi.org/10.1080/15434300701375923.
Fidalgo, AM, Alavi, SM, Amirian, SMR. (2014). Strategies for testing and practical significance in detecting DIF with logistic regression models. *Language Testing*, *31*(4), 433–451. https://doi.org/10.1177/0265532214526748.
Field, J. (2005). The cognitive validity of the lecture-based question in the IELTS listening paper. *IELTS Research Reports*, *9*, 17–65 Retrieved on 15 Oct 2015 from http://www.ielts.org.
Field, A (2009). *Discovering statistics using SPSS*. Los angeles: Sage Publications.
George, AC, & Robitzsch, A. (2014). Multiple group cognitive diagnosis models, with an emphasis on differential item functioning. *Psychological Test and Assessment Modeling*, *56*(4), 405–432.
Geranpayeh, A, & Kunnan, AJ. (2007). Differential item functioning in terms of age in the certificate in advanced English examination. *Language Assessment Quarterly*, *4*(2), 190–222. https://doi.org/10.1080/15434300701375758.
Guilera, G, Gómez-Benito, J, Hidalgo, MD, Sánchez-Meca, J. (2013). Type I error and statistical power of the Mantel-Haenszel procedure for detecting DIF: a meta-analysis. *Psychological Review*, *18*(4), 553–571. https://doi.org/10.1037/a0034306.
IELTS Handbook. (2007). Retrieved on 5 Sept 2015 from http://www.bing.com/search?q=IELTS+handbook+2007.
Harding, L. (2011). Accent, listening assessment and the potential for a shared-L1 advantage: a DIF perspective. *Language Testing*, *29*(2), 163–180 55 (2), 79–94. https://doi.org/10.1037/h0056564.
Harding, L, Alderson, JC, Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: elaborating on diagnostic principles. *Language Testing*, *32*(3), 317–336. https://doi.org/10.1177/0265532214564505.
Hooper, D, Coughlan, J, Mullen, M. (2008). Structural equation modelling: guidelines for determining model fit. *Electroninc Journal of Business Research Methods*, *6*(1), 53–60.
Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnosis modeling: applying Wald test to investigate DIF for DINA model. Journal of Educational Measurement, 51 *(1)*, 98–125. doi:https://doi.org/10.1111/jedm.12036.
In'nami, Y, & Koizumi, R. (2011). Structural equation modelling in language testing and learning research: a review. *Language Assessment Quarterly*, *8*(3), 250–273. https://doi.org/10.1080/15434303.2011.582203.
Jakeman, V, & McDowell, C (2004). *Step up to IELTS*. Cambridge: Cambridge University.
Jakeman, V, & McDowell, C (2006). *Action plan for IELTS*. Cambridge: Cambridge University.
Junker, BW, & 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. (2013). Validating the interpretation and use of test scores. *Journal of Educational Measurement*, *50*(1), 13–14. https://doi.org/10.1111/jedm.12000.
Kane, MT. (2016). Validity as the evaluation of the claims based on test scores. *Assessment in Education: Principles, Policy, & Practice*, *23*(2), 309–311. https://doi.org/10.1080/0969594x.
Khine, MS (2013). *Application of structural equation modeling in educational research and practice*. Rotterdam: Sense Publishers.
Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, *18*(1), 89–114. https://doi.org/10.1177/026553220101800104.
Kim, YH, & Jang, EE. (2009). Differential functioning of reading subskills on the OSSLT for L1 and ELL students: a multidimensionality model-based DBF/DIF approach. *Language Learning*, *59*(4), 825–865. https://doi.org/10.1111/j.1467-9922.2009.00527.
Kimura, H. (2016). Foreign language listening anxiety: a self-presentational view. *International Journal of Listening*, *00*, 1–21. https://doi.org/10.1080/10904018.2016.1222909.

KÖk, I. (2017). Relationship between listening comprehension strategy use and listening comprehension proficiency. *International journal of Listening*, *0*, 1–17. https://doi.org/10.1080/10904018.2016.1276457.
Kunnan, AJ. (1994). Modeling relationships among some test-taker characteristics and performance on EFL tests: an approach to construct validation. *Language Testing*, *11*(3). https://doi.org/10.1177/026553229401100301.
Kunnan, AJ. (1998). An introduction to structural equation modeling for language assessment. *Language Testing*, *15*(3), 295–332. https://doi.org/10.1177/026553229801500302.
Le, L. (2006). Analysis of differential item functioning. *Paper Prepared for the Annual Meetings of the American Educational Research Association in San Francisco*, 7–11.
Li, FM (2008). A modified higher-order DINA model for detecting differential item functioning and differential attribute functioning. In *Unpublished doctoral dissertation*. Athen: University of Georgia.
Li, H, & Suen, HK. (2013). Detecting native language group differences at the sub-skills level of reading: a differential skill functioning approach. *Language Testing*, *30*, 273–298. https://doi.org/10.1177/0265532212459031.
Li, X, & Wang, WC. (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/pd.
London Teacher Training College (2005). *IELTS training course*. London: London Teacher Training College.
Mantel, N, & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719–748. https://doi.org/10.1093/jnci/22.4.719.
McCarter, S (2006). *Tips for IELTS: a must-have book for all IELTS candidates*. Oxford: Macmillan.
Messick, S. (1974). The standard problem: meaning and values in measurement and evaluation. *American Psychologist*, *2*(2). https://doi.org/10.1002/j.2333-8504.1974.tb01034.x/pdf.
Messick, S. (1986). The once and future issues of validity: assessing the meaning and consequences of measurement. *American Psychologist*, *2*(12), 1–24. https://doi.org/10.1002/j.2330-8516.1986.tb00185.x/pdf (1974, 1986, 1995, 1996).
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.
Messick, S. (1996). Validity and wash back in language testing. *Language Testing*, *13*(1), 241–256.
Michaelides, MP. (2008). An illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. *Practical Assessment, Research and Evaluation*, *13*(7).
Monahan, PO, & Ankenmann, RD. (2005). Effect of unequal variances in proficiency distributions on type-I error of the Mantel-Haenszel chi-square test for differential item functioning. *Journal of Educational Measurement*, *42*, 101–131. https://doi.org/10.1177/014662169301700401.
Monahan, PO, & Ankenmann, RD. (2010). Alternative matching scores to control type I error of the Mantel-Haenszel procedure for DIF in dichotomously scored items conforming to 3PL IRT and nonparametric 4PBCB models. *Applied Psychological Measurement*, *34*, 193–210. https://doi.org/10.1177/0146621609359283.
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*.
Newton, PE, & Baird, JA. (2016). The great validity debate. *Assessment in education: principles, policy & practice*, *23*(2), 173–177. https://doi.org/10.1080/0969594x.1172871.
Newton, PE, & Shaw, SD. (2015). Disagreement over the best way to use the world validity and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 281–283. https://doi.org/10.1080/0969594X.2016.1141750.
Ockey, G, & Choi, I. (2015). Structural equation modeling reporting practices for language assessment. *Language Assessment Quarterly*, *12*(3), 305–319. https://doi.org/10.1080/15434303.2015.1050101.
Pae, T. (2004). DIF for examinees with different academic backgrounds. *Language Testing*, *21*, 53–73. https://doi.org/10.1191/0265532204lt274oa.
Pae, T. (2012). Causes of gender DIF on an EFL language test: a multiple-data driven analysis over nine years. *Language Testing*, *29*(4), 533–554.
Phakiti, A. (2008). Strategic competence as a fourth-order factor model: a structural equation modeling. *Language Assessment Quarterly*, *5*(1), 20–42.
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), 3–7.
Rezaee, A, & Shabani, E. (2010). Gender differential item functioning analysis of University of Tehran English Proficiency Test. *Pazhuheshe- Zabanhaye Khareji*, *56*, 89–108.
Rogers, WT, & Yang, P. (1996). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment*, *12*(3), 247–259.
Roussel, S, Gruson, B, Galan, JP. (2017). What types of training improve learners' performances in second language listening comprehension? *International Journal of Listening*, *00*, 1–14. https://doi.org/10.1080/10904018.2017.1331133.
Sawaki, Y (2012). *Factor analysis. The Encyclopedia of Applied Linguistics*. Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal0407.
Sawaki, Y, Stricker, LJ, Oranje, AH. (2009). Factor structure of the TEOFL internet- based test. *Language Testing*, *26*(1), 005–030.
Schmitt, T. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psycho-educational Assessment*, *29*(4). https://doi.org/10.1177/0734282911406653.
Schoonen, R. (2005). Generalizability of writing scores: a structural equation modeling. *Language Testing*, *22*(1), 1–30.
Sireci, SG. (2017). Interview with Stephen G. Sireci on validity. *Journal of Measurement and Evaluation in Education and. Psychology*, *8*(1), 158–168.
Song, MY. (2008). Do divisible sub-skills exist in second language comprehension? A structural equation modeling approach. *Language Testing*, *25*(4), 435–464.

Song, X, Cheng, L, Klinger, D. (2015). DIF investigations across groups of gender and academic background in a large scale high-stakes language test. *Papers in Language Testing and Assessment*, *4*(1), 97–124.
Su, YH, & Wang, WC. (2005). Efficiency of the Mantel, generalized Mantel-Haenzel, and logistic discriminant function analysis methods in detecting for polytomous items. *Applied Measurement in Education*, *18*(4), 313–350. https://doi.org/10.1207/s15324818ame1804.
Tatsuoka, K, Linn, R, Tatsuoka, M, Yamamoto, K. (1988). Differential item functioning resulting from the use of different solution strategies. *Journal of Educational Measurement*, *25*(4), 301–319. https://doi.org/10.1111/j.1745-3984.1988.tb00310.x.
Templin, J, & Henson, RA. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305.
de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7.
de la Torre, J, & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353. https://doi.org/10.1007/BF02295640.
Vandergrift, L. (1997). The Cinderella of communication strategies: reception strategies in interactive listening. *The Modern Language Journal*, *81*(4).
Vandergrift, L. (2006). Second language listening: listening ability or language proficiency? *The Modern Language Journal*, *90*(1).
Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, *40*, 191–210.
Winke, P, & Lim, H. (2014). The effects of testwiseness and test-taking anxiety on L2 listening test performance: a visual (eye-tracking) and attentional investigation. *IELTS Research Reports*, *3*, 5–6.
Wright, SP. (1992). Adjusted P-values for simultaneous inference. *Biometrics*, *48*(1), 1005–1013.
Zhang, W (2006). *Detecting differential item functioning using the DINA model, unpublished doctoral dissertation*. Greensboro: University of North Carolina.
Zhang, L. (2015). Recent research into IELTS reading and listening assessment by Linda Taylor and Cyril Weir (Eds.) *Language Assessment Quarterly*, *12*, 234–238. https://doi.org/10.1080/15434303.2014.1003218.
Zumbo, BD. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing*, *20*(2), 136–147.
Zumbo, BD. (2007). Three generations of DIF analyses: considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233.