Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 13, No. 1, March 2023
ISSN 2476-5880

Validation of a Language Center Placement Test: Differential Item Functioning

Niloufar Shahmirzadi[1]*

Received: 9 April 2022 Accepted: 1 July 2022

Abstract
The documentation of test takers' achievements has been accomplished through large-scale assessments to find general information about students' language ability. To remove subjectivity, Cognitive Diagnostic Assessment (CDA) has recently played a crucial role in perceiving candidates' latent attribute patterns to find multi-diagnostic information rather than single proficiency classification. However, there are some gaps in the literature about in detail investigation of test takers' listening comprehension language ability in responding to placement test items of a public English language center. The present study aims to validate an English placement test at a language center through a retrofitting process. In an exploratory mixed-method design, 449 participants from the same language center, including 274 females and 175 males, were selected. The performance of randomly selected participants in a language center placement test was analyzed by applying the GDINA model from R-studio packages, to detect Differential Item Functioning (DIF). Results of the study revealed DIF in some items since there is some bias in test items. The implication of this study is to provide meaningful interpretations of respondents' attributes and improve teaching and learning by finding the strengths and weaknesses of candidates. For this purpose, the findings derived from the result of the study can raise the awareness of test developers in preparing unbiased items for the placement test, and at the same time, assist test-takers to become more critical of their English language achievements. It is also helpful for materials developers to become aware of developing materials free from bias.

Keywords: Cognitive Diagnostic Assessment; Differential Item Functioning; listening comprehension; placement test

1. Introduction

The term "diagnosis" derives from the Greek word "*diagignóskein,*" which means "to know precisely, to decide, and to agree upon" respectively (Fisseni, 2004, p. 4). Current attempts have been made to apply theories of Cognitive Diagnostic Assessment (CDA) in a variety of fields (Alderson, 2005; Jang, 2009; Kunnan & Jang, 2009); more specifically, to classify test takers based on the patterns acquired. Technically, these patterns are defined in terms of mastery/non-mastery of an attribute, ability, skill, sub-skill, or knowledge. CDA also aims to diagnose learning difficulties in finer-grained details (Embretson, 1998; Leighton &

[1] Department of Foreign Languages, Tehran Central Branch, Islamic Azad University, Tehran, Iran.
niloufar_shahmirzadi83@yahoo.com

 ISSN 2476-5880

Gierl, 2007; Nichols, 1994) to depict a vivid link between learners' latent skills' mastery profiles or mastery of attributes in test items and characteristics of test items.

That is to say, the goals of CDA are showing the relationship between test items and cognitive skills (Jang, 2005), generating and analyzing items according to the cognitive models (Embretson, 1998), and most importantly removing the pitfalls of Item Response Theory (IRT) or Classical Test Theory (CTT). Moreover, in IRT and CTT merely the final outcome is available to report the ability, attribute or skill mastery/non-mastery of each test taker. Also, in IRT or CTT, it is not clear which test item correctly responded to, is learned or mastered by a test taker, or test takers may reply by chance.

To remove these pitfalls, CDA is in stark contrast with CTT and IRT since it is possible to remove the aforementioned problems of CTT and IRT by developing a Q-matrix in CDA (Ravand, 2019, p. 79). It is worth noting that Ravand and Baghaei (2019) provide three classifications in CDA research. The first category designs tests with diagnostic purposes in advance, and then constructs a Q-matrix for each test item. Q-matrix depicts which attribute or skill is going to be matched by per item (Tatsouka, 1983). Here, the obtained results would be highly valuable to depict the strengths and weaknesses. The second category extracts information about the tests through a retrofitting study. Moreover, data collection is mostly related to existing high-stakes tests. And, the third category argues the model selection and fit of the model through the application of different Cognitive Diagnostic Models.

Here, CDA depicts the attributes which are used in the test items. Then by the application of Q-matrices, which have been developed by the researcher, the process of test item analysis through R software, GDINA package, will continue. In CDA-like confirmatory factor analysis models, latent traits or attributes are explained primarily through Q-matrix. This way of explanation in confirmatory model is based on the selection of a model which is in line with the relevant theories.

Q-matrices are defined in a table in terms of the number of all test items in rows and attributes derived from the test in columns. Typically, more than one attribute is derived from per item. That is to say in contrast to IRT or CTT, in CDA attributes like understanding vocabulary and inferencing might be found in each test item. By this, Q-matrix construction considers as an aid to reveal skill mastery profiles of test takers' latent behaviors or latent ability in practice (Tatsuoka, 1983, 1990). However, in developing Q-matrix care should be taken since the selection of specific attributes and the number of attributes would be highly important (Ravand & Baghaei, 2019). CDA assumes as a powerful tool for the purpose of discovering information in detail, on learners' strengths and weaknesses in order to find whether test takers could successfully complete the task (Birenbaum, et al., 1993, p. 443). That is to say, whether test takers learned or mastered test items.

For the present study, multidimensional measurement of listening comprehension was conducted to entail a more detailed profile of learners' understanding. Here, attributes are evidence in assessing learners' mastery/non-mastery of listening comprehension test items. The product of CDA is also in line with Evidence Centered Design (ECD). ECD facilitates to design task types, determine content areas, and address unfamiliar structure. Here, the aim is to elicit learners' latent language ability structure. It helps to specify learning theories that determine the underlying diagnostic assessment system. Through ECD cognitive diagnostic assessment,

 ISSN 2476-5880

designers can determine which attributes as pieces of information are helpful to reveal the statistical models (Anderson, 2003). These attributes are helpful in data coding for statistical analyses. As a result, they can show the extent to which test takers engage in the cognitive process when responding to cognitive diagnostic tasks. In simple terms, the underlying knowledge of respondents can be analyzed by a robust statistical analysis. This is a substantive aspect of construct validity to depict unobserved abilities of test takers from observed test scores, and to assess test content in terms of measurement traits. According to de la Torre and Chiu (2016), the relationship between an attribute or a skill and an item could be meticulously measured. Here, the primary aim of measurement trait is to provide meaningful interpretation for test scores. This would be possible by validating the content of the instrument.

However, due to major drawbacks in language testing research methodology, there are no general conclusions to explain whether the improvement in the performance of the current study was due to a better understanding of the test items, or the additional information in the questions resulted in responding correctly. Moreover, it is not clear whether because of bias in some test items, some candidates with different background knowledge could answer correctly whereas others could not. These cases may result in unclear comprehension of test takers' cognitive traits, which necessitate examining the validity of test items.

Thus, in the present study attempts had been made to investigate the real ability of test takers in listening comprehension test items of a language center placement test. The rationale to choose this skill is that mostly the focus of placement tests is on vocabulary and grammar rather than comprehension skills like listening comprehension. Here, attempts had been made to ensure test materials security in general and test formats and item types in particular. The materials are mostly used to determine the appropriate level for each test taker. As a result, the researcher considered listening comprehension, as one of the comprehension skills, in the present study. Test items were analyzed through CDA to detect the validity of the placement test administered in a public institute.

1.1.Testing Problem Encountered

Language testing is aimed at gathering useful evidence to show the communicative skills of test takers efficiently and affordably. This notion was primarily proposed in terms of validity by Messick (1989). Therefore, the meaning of the scores and the nature of the scores reveal test takers' latent knowledge. Therefore, Messick assumes test score inference is "an integrative evaluative judgment of the degree to which empirical evidence and theoretical rationale support the inferences" (p. 13). Here, making inferences requires finding abilities of candidates beyond the tests. With pure justification of cognitive matters, test developers can take advantage of underlying latent constructs. These embedded constructs reveal the abilities of test takers. In practice, through the application of recent measurement techniques, test developers can analyze test takers' performance. However, there seems a gap in the literature with cognitive diagnostic objectives to validate listening comprehension skill; more specifically, in the literature for a language center placement test to validate construct underrepresentation.

To remove the problem of validity, the number of sufficient items, the number of required attributes, and selection of the authentic texts with appropriate level of difficulty should be taken into consideration in studies. This can be measured a priori in cognitive

diagnostic assessment. Interestingly, in CDA it is possible to ensure whether tests examined fairness and equality through further analysis of Differential Item Functioning (DIF). DIF is defined as different group memberships' probabilities of replying to a set of test items. If DIF exists, test takers cannot take advantage of an equal chance in dealing with test items. That is, the probability of choosing the correct response in each test item with the same attributes is different per individual.

DIF studies have generally been accepted as a necessary standard for test validity in the broader field of testing and assessment. It has exclusively been taken into consideration in the context of English as a Foreign Language (EFL) testing (Park, 2008; Ryan & Bachman, 1992). DIF occurs when the probability of answering an item correctly (controlling for the ability) is different across groups comprising both females and males. This probability of discrepancy shed light on the primary demographic characteristics like gender, age, cultural and academic backgrounds, which are usually targeted for DIF investigation to enhance test quality. Zhang and Shanshan (2011) believe that gender and the background of the test takers affect the DIF studies. The contemporary procedures of detecting DIF are basically implemented under different statistical approaches rather than using merely the most robust approach. As for CDA, DIF studies occur to detect slipping and guessing parameters. Ketabi, et al. (2021) pointed out that adding and removing attributes unscientifically may result in the misclassification of attributes, which underestimates the goal of CDA. These attributes are analyzed and compared with regard to test takers' responses, which are matched on attribute profiles. Angoff (1993, p. 19), in his review of long-standing Educational Testing Service DIF work, notes that "it has been reported by test developers that they are often confronted by DIF results, and no amount of deliberation seems to help explain why some perfectly reasonable items have large DIF values." Zumbo (2007) justifies that this is a matter of neglecting equity in the process of test development, which is rooted in a dearth of knowledge in language testing. This is the matter which many language centers, particularly the high-ranked language center under study suffered from a poor placement test.

2. Review of Literature

Messick (1989) asserts the importance of perceiving students' mental process to answer test items, and then to resolve cognitive weakness in the process of assessment, as the core features of construct validity theory. Messick also notes "the importance of understanding students' mental or psychological process, as opposed to content based behavior" (p. 26). This approach of construct representation attempts "to identify theoretical mechanisms underlying task performance from component of mental process" (Embretson, 1983, pp. 42-45). Here, observing mental process aims to reveal the hidden self-knowledge of test takers implicated in the test performance.

Borsboom, et al. (2004) consider scientific endeavors to justify the interpretation of cognitive behavior in test scores pursuing a rigorous program on validation. This interpretation focuses on students' mental behavior for enhancing their opportunity to learn through the application of different models in CDA (Chen & Chen, 2016; Lee & Sawaki, 2009). In practice, diagnostic assessment is "a systematic process that seeks to obtain specific information about psychological characteristics of a person by using a variety of methods" (Kubinger, 2006, p.

4). With the advent of CDA, it has been widely used to investigate the cognitive model, and the diagnostic information of comprehension tests such as reading and listening comprehension. With regard to the importance of reading comprehension, recently the majority of studies (e.g., Hemati & Baghaei, 2020; Ketabi, et al., 2021; Ranjbaran & Alavi, 2017; Roohani Tonekaboni, et al., 2021; Shahmirzadi, et al., 2020 a, b; Tabatabaee-Yazdi, et al., 2021) analyzed reading comprehension test items in both formative and summative tests. Moreover, Kunnan, et al. (2022) discussed some challenges of developing and administering a test with placement and diagnostic purposes. As for the present study, listening comprhension test items, as the necessary comprehension skill (e.g., Jang, 2009; Li, et al., 2015; Rupp, et al., 2006) were analyzed. In practice, it is necessary to introduce some models in CDA, and then it is crucial to take into account the selection of appropriate cognitive diagnostic model, which is required to reflect the underlying knowledge of test takers. There are some special cognitive diagnostic models, including General Multicomponent Latent Trait Model (GLTM) (Embretson, 1984) as a successful example of the first attempts in cognitive diagnostic modeling, Tatsuoka's Rule-Space Method (RSM) (Tatsuoka, 1985; 1995) as a clear progression in the field of CDM, Log-linear Cognitive Diagnostic Model (LCDM) (Henson, et al., 2009), the Fusion Model (Hartz, 2002), General Diagnostic Model (GDM), Deterministic-Input, Noisy-and-Gate Model (DINA) (von Davier, 2011), and the Generalized version of the DINA model (GDINA) (de la Torre, 2011), which categorize in the most advanced family of the parametric CDM methods. Among all, the package of GDINA (the Generalized Deterministic Input, Noisy, and Gate Model) model is the most user-friendly model, especially in case of specifying model fit indices. For the present research, GDINA model was used for analyzing data. It is a model which is available in R-studio software. To run GDINA model, some programs should be fed to the R-studio, and then run the model. The details of GDINA outputs are available in Result section of the present study.

In addition, there seems to be a dearth of research on non-diagnostic assessment contexts, which is known as retrofitting in CDA. Through retrofitting, it is possible to extract diagnostic information from pre-existing high-stakes tests. During recent decades, research on retrofitting has been gradually enhancing. Lee and Sawaki (2009, p. 174) claim that "retrofitting efforts could serve as an important step in advancing cognitive diagnostic assessment research before delving into the process of designing a new diagnostic test." In the meanwhile, Javidanmehr and Anani Sarab (2019), Kim (2015), and Mirzaei, et al. (2020) believe that retrofitting CDA studies have been used in a wide variety of fields of non-diagnostic high-stakes testing in CDMs. Applying retrofitting also has some advantages, including both saving huge amounts of time and budget to develop a diagnostic test (Lee & Sawaki, 2009), and providing detailed and fine-grained information on applicants rather than merely reporting applicants' total scores. Of course, care should be taken since applying retrofitting in unidimensional CDMs may cause serious consequences (Gorin, 2009; Tatsuoka, 2009).

2.1.Purpose of the Study

As for the present study, it would also benefit the researchers and teachers to gain insights on identifying sub-skills or attributes in the English language placement test. That is to say, teachers can become more aware of their test takers' needs, and how teachers can remove students' weakness. In compensatory function, low abilities of test takers in one attribute can

be compensated by high abilities in another attribute to answer a test item. As a result, removing listening comprehension pitfalls as one of the comprehension skills was taken into consideration. In other words, the current study was an attempt to analyze listening comprehension test items of a public language center placement test. To do so, the following research question posited.

Is there any DIF in listening comprehension test items under CDA?

3. Methodology

To measure listening skills, the sequential exploratory mixed method design was adopted. According to Jang (2005), in cognitive diagnosis assessment many perspectives need to be considered in order to have ample evidence concerning the validity of the obtained results from the diagnostic inferences. In this research, the priority was given to qualitative and then quantitative data collection and analysis. In the initial qualitative exploration, content analysis of listening comprehension test items through conducting a think-aloud verbal protocol analysis was fulfilled to develop a Q-matrix. Then, the researcher estimated the DIF for each test item. Noteworthy, as it is common in the sequential exploratory mixed-method design, the participants in the quantitative study were larger, and they were not the same individuals who provided qualitative data.

3.1. Participants

Collected data for the current study were narrowed by the population of English language learners at a high-ranking public language center in Tehran, attending placement tests over time. The researcher analyzed listening comprehension of 449 simple random sampled test takers who were intending to pursue English as a foreign language at the center from A_1 to C_2. The candidates were between the age range of 18 to 40 comprising both 175 males and 274 females who were majoring at university in different fields. To analyze the participants' listening comprehension and their skill mastery/non-mastery profiles, it is crucial that levels and topics of the developed test were in line with the general English course books.

3.2. Data Collection

To analyze listening comprehension of the placement test, a listening section was selected. It was a four-option formative test held four times in a year. Candidates who pursue their English language were required to sit for the test in 60 minutes in total before starting each semester in an English language center. They admitted to the programs from A_1 to C_2 based on the Common European Framework of Reference (CEFR). The test was composed of 10 listening items. The tests were administered four times in a year. On each testing occasion, the subject group, and other students took the test simultaneously at the exam venue.

In listening comprehension section, attempts had been made to examine skills mastery/non-mastery profiles of test takers in 5 attributes as follows: understanding vocabulary (Wolfgramm, et al., 2016), understanding details (Sawaki, et al., 2009), inferencing (Hildyard & Olson, 1978; Wagner, 2004), paraphrasing (Wagner, 2004), and differentiating main ideas (Yeldham, 2016).

 ISSN 2476-5880

3.3 Q-Matrix Development

To encode the relationship that exists between diagnostic assessment items and latent variables, Q-matrix is used to classify items. In practice, to develop a Q-matrix, two females and two males among English language learners who were studying at either B2 or C1 levels were randomly recruited by the researcher to participate in a think-aloud verbal protocol analysis. Ravand and Baghaei (2019) believed that think-aloud protocol reflects the underlying processes at the time of taking the test. Participants were supposed to read the questions and express their attitudes regarding listening sub-skills or attributes such as understanding vocabulary, understanding details, inferencing, paraphrasing, and differentiating main ideas. In doing so, first the researcher explained the stages of conducting this phase of the study. That is, she asked the participants to do the task. They talked about what they thought while listening and responding. Participants were supposed to do the test and verbalize their thoughts immediately so as not to forget their natural thoughts. Then, a panel of experienced English language teachers, who had been teaching for almost 10 to 13 years at different language centers and the language center under study, was invited to examine the extent to which each listening attribute contributed in answering to each test item. They hold an MA degree in Teaching English as a Foreign Language (TEFL). As a result, the refined Q-matrix was constructed for the listening comprehension section. According to Sawaki, et al. (2009, p. 195), the following steps are a general overview of Q-matrix development in comprehension test items.

A. Review of test documents, literature, and brainstorming possible approaches,

B. Task analysis of test items,

C. Defining skills and item coding,

D. Empirical analysis of the examinees' performance data (GDINA model).

As for the current study, the attributes of the finalized Q-matrix for listening comprehenssion test items depict in Table 1.

Table 1

Listening Comprehension Q-Matrix for English Language Placement Test

Items	Understanding vocabulary	Understanding details	Inferencing	Paraphrasing	Differentiating main ideas
Q1	1	1	1	0	0
Q2	0	0	0	0	1
Q3	1	0	0	1	1

ISSN 2476-5880

Q4	1	0	0	1	0
Q5	0	1	1	0	1
Q6	1	1	0	1	0
Q7	0	1	1	0	1
Q8	1	0	0	1	1
Q9	0	1	1	1	0
Q10	1	0	1	1	0

3. Results

Through the application of R-studio package, data were fed. Then, from this package GDINA model, which relaxes the assumption of equal probability of success, was applied to test model fit indices, and detect DIF. To accomplish the model fit indices, Akaike Information Criterion (AIC) (Akaike, 1974), Bayesian Information Criterion (BIC) (Schwarzer, 1978), and log likelihood (loglike) were measured. According to Li, et al. (2015), and Rupp, et al. (2010), the smallest value for the AIC and BIC indices were preferred to ensure the best fitting model in any cognitive diagnostic statistical model.

Table 2 describes AIC and BIC in the relative fit index. Results present that AIC=3846, and BIC=4216 carried low values, and the lowest value flagged for AIC. That is to say, GDINA model was the best fitting model in the current study.

Table 2

Gender -2Log Likelihood Indices in Listening Comprehension English Language Placement Test

Listening comprehension GDINA model	AIC	BIC	-2Loglike
	3846	4216	-1833.14

In CDM, there are a number of fit statistic, which report on the model fit indices, such as Mx^2, MADcor, $MADQ^3$, SRMSR, and RMSEA. Mx^2 is the global model fit; MADcor statistic estimates the observed, and model-predicted item correlations (DiBello, et al., 2007). $MADQ^3$ statistic calculates the average of the pairwise correlation of these residuals (Yen, 1984). SRMSR estimates the average difference between the observed and predicted correlation matrices. And, RMSEA shows the fit of the items.

To observe the fit of the model, some indices were estimated as shown in Tables 3 and 4. In detail, the estimated mean of items (RMSEA<0.05) is statistically significant (RMSEA=0.039) for each 10 test items.

Table 3

Gender Item Fit Statistic in Listening Comprehension English Language Placement Test

Listening comprehension items	RMSEA
Q1	0.005
Q2	0.014
Q3	0.011
Q4	0.010

Q5	0.000
Q6	0.015
Q7	0.023
Q8	0.007
Q9	0.012
Q10	0.019

Table 4

Gender Model Fit Indices in English Language Listening Placement Test

Goodness of fit indices	Estimate
Mx^2	3.17549173 (p=1.0000000)
MADcor	0.02326410
SRMSR	0.03170291
$MADQ^3$	0.06019136
$MADaQ^3$	0.05153910
Mean of items RMSEA	0.039

As Table 4 depicts, some statistical model for fit indices were as follows:

Mx^2=3.175 (p>.05), MAD=0.023, SRMSR=0.031, $MADQ^3$=0.060, and RMSEA=0.039 (p<.05). Results obtained revealed that GDINA was a precisely adopted model in this study.

Items and attributes parameters were estimated through GDINA model for gender differences. Participants' skill mastery probabilities in the listening comprehension test were shown (Table 5). The differences presented which group had a higher/lower chance to master in the listening comprehension skills.

Table 5

Skill Probabilities based on Gender Differences

Skills Genders	Understanding vocabulary	Understanding details	Inferencing	Paraphrasing	Differentiating main ideas
Male	0.7347	0.4121	0.4674	0.5550	0.5687
Female	0.2653	0.5879	0.5326	0.4450	0.4313
Difference	0.4694	-0.1758	-0.0652	0.1100	0.1374

For males understanding vocabulary, paraphrasing, and differentiating main ideas, and for females understanding details and inferencing had lower probabilities in comparison to other skills. Thus, except for understanding details and inferencing, all the other skills for males were uniformly higher than females.

To detect DIF in the last stage of the study, it is vital to show that a set of sub-skills is equal to some values (Hou, et al., 2014). Armstrong (2014) asserts that the adjusted p-value was improved by the Bonferroni method in multiple groups DIF detection. The outputs of DIF detection were provided (Table 6).

 ISSN 2476-5880

Table 6

Gender DIF Detection in English Language Listening Placement Test

Items	Wald statistic	DF	P-value	Adjusted P-value
1	125.9158	8	0.0000	0.0000**
2	0.0060	2	0.9970	1.0000*
3	7.3735	8	0.4969	1.0000*
4	19.7030	8	0.0115	0.1152*
5	15.7282	8	0.0464	0.4644*
6	12.6845	4	0.0129	0.1292*
7	10.5882	8	0.2261	0.0000**
8	47.0001	8	0.0000	0.0000**
9	40.6363	16	0.0006	0.0063**
10	10.9760	4	0.0268	0.2684*

Note: adjusted p-values are based on the Bonferroni correction.
Note: Effect Size Evaluation is based on Unsigned Area (UA)
** Non-Significant >.05*
*** Large 0.00 - 0.088*

In Table 6, adjusted p-value suspected large DIF in items 1, 7, 8, and 9. The other items were also non-significant in case of DIF detection. Results of the study showed that females could take advantage of inferential questions and listening for details since they were more analytic. On the other hand, males were mostly outperformed in vocabulary, paraphrasing, and differentiating main ideas.

5. Discussion

Regarding the influence of cognitive psychology on educational measurement, Snow and Lohman (1989) explicitly propose that "cognitive psychology is a central concern for educational measurement" (p. 265). In other words, assessment of test takers' learning in practice would appear within the realm of cognitive psychology matrix. To reconsider the research question, in cognitive psychology, cognitive diagnostic can complete the assessment. That is to say, through developing quality matrix attributes, which measured each item, details of per item can be depicted. In the meanwhile, the acceptable fit of the data may appear, which result in the assignment of correct identification of attributes in CDM model. As for the present study, these attributes which were selected based on their importance in existing literature were understanding vocabulary, understanding details, differentiating, inferencing, paraphrasing, and differentiating main ideas. These attributes helped in understanding longer chunks of discourse in listening comprehension. Here, some details derived from these attributes revealed more information about test takers' pitfalls. This may help instructors to focus more on students' weaknesses in the classroom. More specifically, there are test takers who cannot take advantage of one attribute such as paraphrasing to answer another test item like inferencing. Grab (2009) also mentioned that inferencing was identified as one of the difficult attributes since it needs higher level of information processing from the text.

Moreover, Nicolas (1994) outlines a substantive–driven test development approach in the CDA. It is included the importance of "substantive theory construction, design selection, test administration, response scoring, and design revision" (p. 587) to improve quality assessment. Otherwise, some sorts of bias, or statistically differential item functioning may appear in test items. Gómez-Benito, et al. (2018) emphasize that within the last two decades, DIF has been considered as a possible threat to test validity. As for the present study, some items suffer from DIF, and different performance of both genders appears in the findings, which put the quality of the placement test under investigation. In the meanwhile, many scholars refer to negative washback effect for language teaching, because many language teachers may take into account some weaknesses of test takers, which derived from some studies and neglect thorough teaching of listening comprehension attributes (Chen & Chen, 2016; Lee & Sawaki, 2009; Ravand, 2016). As a result, test method effect is impossible to eliminate (Bachman, 1990).

In addition, in order to achieve a thorough understanding of the DIF in test items, detect and remove them here, the researcher tried to depict a deeper understanding by using a retrofitting procedure. To do so, Ravand and Baghaei (2019) propose that the selection of specified attributes is more informative because it is possible to find the depth of candidates' perception. In the last analysis of DIF, the results revealed that some items were suspected to DIF whereas the others were not statistically significant. Thus, results of statistical analysis showed that there is no considerable contribution of the language center placement test to the listening comprehension since test developers were not clearly aware of detailed blueprints in the process of item development and validation.

6. Conclusions and Implications

Implementing democratic norms in language testing have been a complex task for test developers to examine structural validity with the aim of observing structural equalities, on the one hand. Therefore, this problem may arouse numerous difficulties for test developers. To remove it, adding flexibility and variability to the test is possible so as to be able to build new variable. This may help to better support to test the language with authenticity. On the other hand, most of the supervisors or teachers, who had a contribution to test services, were not expert enough in language testing and assessment. To resolve the problem, it is necessary to take into consideration the elements of listening assessment grids (Council of Europe, 2001), because they would be a valid rubric for designing listening comprehension test items aiming to enhance proficiency. That is, the accuracy of decision with regard to setting a standard is crucial. Here, CEFR rubrics could present skill mastery profiles of test takers qualitatively and quantitatively. The former refers to the effective use of the task, and the latter explains the number of skill mastery profiles while doing a task (Council of Europe, 2001).

Seemingly, in CDMs multiple-choice questions can be constructed to measure higher order of cognitive levels. Most importantly, in these questions distractors can respond to student misconceptions. Pek and Poh (2004) argue that examining students' incorrect answer can provide empirical evidence against specific areas of skill mastery/non-mastery profiles in diagnosis and further instruction. Thus, cognitive diagnostic assessment models are considered as the assessment for learning; in contrast to CTT and IRT since they are used as the assessment of learning.

Thus, it is assumed that the accumulation of evidence would be the ultimate goal of Evidence–Centered Design (ECD) through CDA in collecting ample evidence, localizing listening assessment components in the process of item development. As for the present study, the researcher believes that tests should be relatively fair, and the justice of the test could be open to dispute. As a result of which, researchers are required to expand wider perspectives in a test development with regard to social values. This logic can also be spread with higher level of mental skills and abstractions. In practice, organismic variables, including physical and psychological variables aid to assessment. And, interaction access, which is an aid for test takers to encounter a wide variety of fields like particular cultures or understanding ontology, is important to reconsider in assessment.

Moreover, enhancing pragmatic perspective is the helpful evidence for test developers to apply suitable measurement models, because they can perceive test takers' real world knowledge, ethics, and situation. Psychometric model adopted can also assist in measuring of attributes meticulously.

In the end, an infinite number of test items could be generated properly in language testing with regard to test takers' psychological and cognitive factors. It is also worth paying heed to how these features affect test takers' motivation. It is of paramount importance both to avoid the results of test misuse, and to find the reasons why these tests are invalid for interpretation and use.

Such efforts should be recommended, in a great many other attributes for listening comprehension test items, and another comprehension test that is reading comprehension test items. Of course, attempts should be made to decrease the limitation of subjectivity in a Q-matrix construction to enhance objectivity in the process of test development. It is also crucial to increase the number of participants for verbal protocol analysis. This approach, which to some extent is found it difficult to achieve, may need collaboration with a number of experts in CDA. In the end, it is suggested to collect data from a large variety of population from different institutions to widely compare the result obtained.

Declaration of Conflicting Interests

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

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum. https://doi.org/10.1080/15434300701595637

Anderson, N, J. (2003). Scrolling, clicking, and reading English: Online reading strategies in a second/ foreign language. *The Reading Matrix*, *3(3)*, 1-33. http://www.readingmatrix.com

Angoff, W. (1993). Perspective on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum Associates. https://www.ets.org/research/policy_research_reports/publications/chapter/1993/hrez

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics, 34*(5), 502-508. https://doi.org/10.1111/opo.12131

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

Birenbaum, M., Kelly, A. E., & Tatsuoka, K. K. (1993). Diagnosing knowledge states in algebra using the rule-space model. *Journal of Research in Mathematics Education*, 24, 442-459. https://psycnet.apa.org/doi/10.2307/749153

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071. https://psycnet.apa.org/doi/10.1037/0033-295X.111.4.1061

Chen, H. L., & Chen, J. S. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly*, *13*(3), 218-230. https://doi.org/10.1080/15434303.2016.1210610

Council of Europe. (2001). *Common european framework for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press. https://rm.coe.int/16802fc1bfs

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

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

DiBello, L.V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C.R. Rao, & S. Sinharay (Eds.), *Handbook of statistics*, Vol. 26: *Psychometrics* (pp. 970-1030). Amsterdam: North-Holland Publications. https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1342084

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49*(2), 175-186. https://doi.org/10.1007/BF02294171

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380-396. https://psycnet.apa.org/doi/10.1037/1082-989X.3.3.380

Fisseni, H. J. (2004). *Lehrbuch der psychologischen Diagnostik [Psychological assessment].* Göttingen: Hogrefe.

Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D., & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema, 30*(1), 104-109. https:// doi: 10.7334/psicothema2017.183

Gorin, J. S. (2009). Diagnostic classification models: Are they necessary? Commentary on Rupp and Templin. *Measurement: Interdisciplinary Research and Perspectives, 7*(1), 30–33. https://doi:10.1080/15366360802715387

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*.

Cambridge: Cambridge University Press.

Shahmirzadi, N., Siyyari, M., Marashi, H., & Geramipour, M. (2020a). Selecting the best fit model in CDA: DIF detection in reading comprehension PhD nationwide admission test. *The Journal of Language and Translation*, *10*(3), 1–15. https://ttlt.stb.iau.ir/article_678751.html

Shahmirzadi, N., Siyyari, M., Marashi, H., & Geramipour, M. (2020b). Test fairness analysis in reading comprehension PhD nationwide admission test items under CDA. *Journal of Foreign Languages Research*, *10*(1), 152–165. https://jflr.ut.ac.ir/article_75588.html

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive Abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois at Urbana–Champaign, Urbana, IL. https://www.semanticscholar.org/paper/A-Bayesian-framework-for-the-unified-model-for-with-Hartz/31843907beb32ef44f25780a96a340bdb8e4e2ab

Hemati, S., & Baghaei, P. (2020). A cognitive diagnostic modeling analysis of the English Reading comprehension section of the Iranian national university entrance examination. *International Journal of Language Testing*, *10*(1), 11-32. https://www.ijlt.ir/article_114278.html

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log–linear models with latent variables. *Psychometrika*, *74*(191). https://link.springer.com/article/10.1007/s11336-008-9089-5

Hildyard, A., & Olson, D. (1978). Memory and inference in the comprehension of oral and written discourse. *Discourse Processes, 1*(2), 91-107. https://doi.org/10.1080/01638537809544431

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*(1), 98-125. https://doi.org/10.1111/jedm.12036A

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign. [Available from ProQuest Dissertations and Theses database. (AAT 3182288)]. http://hdl.handle.net/2142/79837

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing*, *26*(1), 31-73. https://doi.org/10.1177%2F0265532208097336

Javidanmehr, Z., & Anani Sarab, M. R. (2019). Retrofitting non-diagnostic reading comprehension assessment: Application of the G-DINA model to a high stakes reading comprehension test. *Language Assessment Quarterly, 16*(3), 294-311. https://doi:10.1080/15434303.2019.1654479

Ketabi, S., Alavi, S. M., & Ravand, H. (2021). Diagnostic test construction: Insights from cognitive diagnostic modeling. *International Journal of Language Testing, 11*(1), 22-35. https://www.ijlt.ir/article_128357.html

Kim, A.Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2),

227–258.https://doi:10.1177/0265532214558457

Kubinger, K. (2006). *Psychologische Diagnostik: Theorie und Praxis psychologischen Diagnostizierens* [Psychological assessment: Theory and practice]. Göttingen: Hogrefe.

Kunnan, A. J., Qin, C. Y., & Zhao, C. G. (2022). Developing a scenario-based English language assessment in an Asian university. *Language Assessment Quarterly, 19*(3), *Published Online.* https://doi.org/10.1080/15434303.2022.2073886

Kunnan, A. J., & Jang, E. E. (2009). Diagnostic feedback in language assessment. In M. Long & C. Doughty (Eds.), *Handbook of second and foreign language teaching* (pp. 610-625). Walden, MA: Wiley-Blackwell.

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

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. New York: Cambridge University Press. https://psycnet.apa.org/doi/10.1017/CBO9780511611186

Li, H., Hunter, C. V., & Lei, P. W. (2015). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing, 33*(3), 391–409. https://doi.org/10.1177/0265532215590848

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

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education/Macmillan.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*(4), 575-603. https://psycnet.apa.org/doi/10.2307/1170588

Park, G. P. (2008). Differential item functioning on an English listening test across gender. *TESOL Quarterly, 42*(1), 115-123. http://www.jstor.org/stable/40264430

Pek, P. K., & Poh, K. L. (2004). A Bayesian tutoring system for Newtonian mechanics: Can it adapt to different learners? *Journal of Educational Computing Research, 31*(3), 281-307.http://dx.doi.org/10.2190/VDAP-K5BX-EJX1-D8QY

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

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

Ravand, H. (2019). Application of a hieratical diagnostic classification model in assessing reading comprehension. In V. Aryadoust, V., & M. Raquel (Eds.), *Quantitative data analysis for language assessment, Volume II* (pp. 77-98). Routlege. https://doi.org/10.4324/9781315187808

Ravand, H., & Baghaei, P. (2019). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing*, *20*(1), 24-56.

https://doi:10.1080/15305058.2019.1588278

Roohani Tonekaboni, F., Ravand, H., & Rezvani, R. (2021). The construction and validation of a q-matrix for a high-stakes reading comprehension test: A G-DINA study. *International Journal of Language Testing*, *11*(1), 58-87. https://www.ijlt.ir/article_128361.html

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, *23*(4), 441-474. https://psycnet.apa.org/doi/10.1191/0265532206lt337oa

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic assessment: Theory, methods, and applications.* New York: Guilford. https://www.routledge.com/Diagnostic-Measurement-Theory-Methods-and-Applications/Rupp-Templin-Henson-Gierl-Chen/p/book/9781606235270

Ryan, E. K., & Bachman, L. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing, 9*(1), 12-29. http://dx.doi.org/10.1177/026553229200900103

Sawaki, Y., Kim, H. J., & Gentile, C. (2009). Q-Matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly*, *6*(3), 190-209. https://doi.org/10.1080/15434300902801917

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464. https://www.jstor.org/stable/2958889

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.*), Educational measurement* (3rd ed.) (pp. 263-331). New York: American Council on Education/Macmillan. https://psycnet.apa.org/record/1989-97348-007

Tabatabaee-Yazdi, M., Motallebzadeh, K., & Baghaei, P. (2021). Mokken scale analysis of an English reading comprehension test. *International Journal of Language Testing, 11*(1), 132-143. https:// www.ijlt.ir/article_130373.html

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345-354. https://www.jstor.org/stable/1434951

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 10*(1), 55-73. https://psycnet.apa.org/doi/10.2307/1164930

Tatsuoka, K. K. (1990). Toward an integration of item–response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Er https://psycnet.apa.org/record/1990-97343-018lbaum

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P.D. Nichols, S.F. Chipman, R.L.Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). Hillsdale, NJ: Lawrence Erlbaum Associates. https://psycnet.apa.org/record/1995-97594-013

Tatsuoka, K. K. (2009*). Cognitive assessment: An introduction to the rule space method.*

Routledge. https://doi: 10.4324/9780203883372

von Davier, M. (2011). *Equivalency of the DINA model and a constrained general diagnostic model.* ETS research report, (pp. 11-37). Princeton, NJ: Educational Testing Service. https://files.eric.ed.gov/fulltext/ED525306.pdf

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

Wolfgramm, C., Suter, N., & Göksel, E. (2016). Examining the role of concentration, vocabulary and self-concept in listening and reading comprehension. *International Journal of Listening*, *30*(1-2), 25-46. https://doi.org/10.1080/10904018.2015.1065746

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

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125-145. https://journals.sagepub.com/doi/10.1177/014662168400800201

Zhang, Y., & Shanshan, H. (2011). Background knowledge and reading comprehension. *In International Conference on Computer Technology and Development, 3rd* (ICCTD 2011). ASME Press.

Zumbo, B. D. (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. https://psycnet.apa.org/doi/10.1080/15434300701375832