**Journal of**
**Language and Translation**
Volume 10, Number 3, 2020, (pp.1-15)

# Selecting the Best Fit Model in Cognitive Diagnostic Assessment: Differential Item Functioning Detection in the Reading Comprehension of the PhD Nationwide Admission Test

**Niloufar Shahmirzadi[1], Masood Siyyari[2*], Hamid Marashi[3], Masoud Geramipour[4]**
[1] PhD Candidate of TEFL, Department of Foreign Languages, Central Tehran Branch, Islamic Azad University, Tehran, Iran
[2] Assistant Professor of Applied Linguistics, Department of Foreign Languages, Tehran Science and Research Branch, Islamic Azad University, Tehran, Iran
[3] Associate Professor of Applied Linguistics, Department of Foreign Languages, Central Tehran Branch, Islamic Azad University, Tehran, Iran
[4] Assistant Professor of Assessment, Kharazmi University, Tehran, Iran

Received: 25 December, 2019 Accepted: 3 March, 2020

**Abstract**
This study was an attemptto provide detailed information of the strengths and weaknesses of test takers' real ability through cognitive diagnostic assessment, and to detect differential item functioning in each test item. The rationale for using CDA was that it estimates an item's discrimination power, whereas classical test theory or item response theory depicts between rather within item multi-dimensionality. To fulfill the purpose of this study, latent attributes are shown in a Q-matrix and 4200 participants who sought to pursue their studies at the PhD level at state universities were randomly selected. The test used for the present research consisted of two different reading passages with 10 multiple-choice items consisting of four options. The data were analyzed with the application of R studio package, GDINA, and DINA models. Item and model fit indices were estimated and the Wald test was run. The result of the study revealed that some items flagged DIF. The study further concluded that CDA can provide pedagogically useful diagnostic information for test designers, teachers, syllabus and materials developers, and policymakers as a proficiency test needs to be valid, reliable, and fair in the context of high-stakes tests so that it improves the knowledge of test takers.

**Keywords:** Cognitive diagnostic assessment; Differential item functioning; Reading comprehension

## INTRODUCTION

According to the Standards for Educational and Psychological Testing (American Educational Research Association (AERA), American Psychological Association, and National Council on Measurementin Education, (American Educational Research Association (AERA), 1999), "validity

*****Corresponding Author's Email: m.siyyari@srbiau.ac.ir

refers to the degree to which evidence and theory support the interpretations of test scores entailed by uses of tests" (p. 9). As a result, validation of psychological tests needs to conceptualize what qualities should be specified in advance to develop a coherent and valid test. This would require identification of valid tests in terms of construct validity.

The earliest idea of construct validity was developed to investigate the universe of collected evidence to define psychological test performance (Cronbach, 1955). Bechtoldt(Bechtoldt, 1951) asserts that "construct validity involves the acceptance of a set of operations to measure underlying abilities" (p. 1245). That is to say, construct validity seeks to reveal the nature of latent criterion when it is not available(Gaylord, 1955).In this case, determining ultimate standards in the process of test validation is necessary. Construct validity occurs when "indirect measures" could be used "to determine and enhance the quality of a test" (Technical Recommendations for Psychological Tests and Diagnostic Techniques) (Techniques, 1954). In construct validity attempts need to be made to

determine how different individuals with strong or poor competency differentiate from each other with regard to interpretation of data. Because the concept of a construct is the reflection of what thetest performance is, which can be used in theinterpretation of a construct. Messick(Messick, 1995)also emphasizes both consequences of score interpretation and test useto enhance construct validity.

In this regard, Snow and Lohman (Snow, 1989) note that through CDA making realistic assumptions about the latent variables that affect performance on items, revealing thepsychological processes that delineate the construct measured by a test, and defining item responses through a substantive psychological theory are possible. Therefore, in CDA it is possible to measure unlimited latent attributes or sub-skills in a fine-grained size to indicate in which specific skills have or have not been mastered. Alderson(Alderson, 2005)believes that the finer the grain size, the more detailed the information might delineate.CDA as a recent mode of assessment also overcomes some limitations in language assessment, comparing to CTT or IRT. However Ravand and Robitzsch (Ravand, & Robitzsch, 2015) believe that"the application of CDA has not widespread enough. Because CDA is relatively new and its theoretical underpinnings have not been explicated thoroughly" (pp. 1-2). For example, there are a considerable number of controversies in terms of the optimal number ofsample sizes, latent traits or attributes, difficulty in interpreting model fit indices, and results.

Historically, CDA research and application originated in educational measurement–reading comprehension skill(Jang, 2005;Kim, 2015; Li, 2011;Ravand, 2016) – and mathematics assessment more specifically (Henson, 2009; Tatsuoka, 1990; Torre, & Douglas, 2004). Through a non-diagnostic framework known as retrofitting in reading comprehension assessment, most CDA studies have been developed and calibratedto analyze test items (Aryadoust, 2011;Chen, & Chen, 2016;Kim, 2015;Li, 2011; Ravand, Barati, & Widhiarso, 2012; Templine, 2014). However, these studies have ignoredto assess DIF in reading comprehension section of PhD nationwide admission test in Iran.For example, some studies have currently been conducted for BA and MA levels with cognitive diagnostic objectives(Baghaei, & Ravand, 2015;Hemmati, 2016; Ranjbaran, 2017; Ravand, 2015).Regarding the importance of CDA in validation of high-stakes tests, attempts need to be made to provide much more evidence on validity of nationwide university entrance examinations, because it could influence on future of test takers. However, in the absence of ample evidence on validity of nationwide university entrance examination in PhD level, the situation even gets worse. Thus, the present study contributes to the Iranian context and the larger applied linguistics community. On the local context, administering valid tests affect the lives of a large number of test takers. However, little empirical evidence exists to support the validity of the high-stakes tests in PhD level as it is not cost effective. Therefore, the present study aims to examine validity of Iranian National University Entrance Examination (INUEE). In a broad con-

text, some problems pertain to the construct irrelevant variances which reside in reading comprehension test items can be estimated.

Therefore, as this area is untouched, in the present study validity of a high-stakes test was addressed with respect to differential item functioning of each test item under CDA. The failure on the test results is a serious consequence for the examinees since they may consider getting into university as a way to succeed. In addition, one year test preparation causes test takers to undertake serious economic problems. Regarding such grave consequences, it is highly critical to provide transparency in validity of test development and use from CDA perspective.Because this test affects the lives of tens of thousands of test takers each year; therefore, every effort is expected to enhance the validity of test.

**Reading Comprehension**

Reading is the most important skill to ensure success in learning (Anderson, 2003). That is to say, having fluency in reading skills lead into mastery of building meaning in second language materials. Jang (Jang, 2005)emphasizes that reading comprehension consists of "complex, covert, mental, and social activities that interact with a number of both human and environmental factors" (p. 3). Hence, advances in perceptions of reading could establish a better ground for test takers with regard to dominant contexts. This may result from having successful processing of language and interaction with the worldknowledge.

Regarding the nature of reading, it is a multifaceted and complicated language skill which is closely related to different variables in the process of comprehension. Although still there has not been a thorough consensus among psychometricians, educationalists, and psychologists to constituent of reading ability, the importance of learningmultiple reading sub-skills has been proved (Alderson, 1990a, 1990b, 2000;Lumley, 1993; Rost, 1993).

As for the importance of learning skills or attributes in reading comprehension, Grabe(Grabe, 1991) proposes some attributes including knowledge of vocabulary and structure, discourse structure, world background knowledge, and metacognitive knowledge and skills monitoring. Alderson(Alderson, 2000) also emphasizes that attributes mastery in reading comprehension are closely intervowen to the text and reader. Salager-Meyer (Salager-Meyer, 1991) and Cooper(Cooper, 1984) believe that poor vocabulary and structural knowledge could neither deduce the meaning of unfamiliar words, nor identify semantic relationships between the lines. In addition, Gao and Rogers (Gao, 2010), and Jang (Jang, 2009)assert the key roles of

some skills such as vocabulary, syntax, extracting explicit information, connecting and synthesizing, and making inferences to enhance reading ability.

In the early twentieth century, some scholars refute any underlying mental processes in reading skill (Watson, 1913) as they follow behavioristic and observable paradigm. But gradually, both perceiving the literal meaning of unauthentic texts, and explicating its meaning through observable sub-skills are emphasized (Clapham, 1996;Langer, 1992). A few years later, the prominent role of mind and cognitionhighlighted. Chomsky (Chomsky, 1957)accentuates the key role of cognitive processing and schema processing in meaning construction. Gough (Gough, 1972) adds the visual decoding and encoding of letters, which obtained by reaching automaticity in component processing (LaBerge, 1974). Goodman (Goodman, 1967) calls "a psycholinguistic guessing game" (p. 135) which demands a hierarchical processing and active engagement of mind to decipher meaning (Anderson, & Pearson, 1988; Carrell, 1988; Goodman, 1967; Smith, 1971). Subsequently, most recent theories highlight a more interactive and discourse processing view in reading comprehension which could compensate for deficiencies in reading process (Grabe, 1988, 2000; Kintsch, 1978; Rumelhart, 1977, 1980; Stanovich, 1980). By this, readers can move towards interaction with the micro-structure and macro-structure of the text in order to construct

meaning based on the underlying information retrieved from the memory. That is justified by

Anderson and Pearson (Anderson, & Pearson, 1988), and Clapham (Clapham, 1996)that background knowledge of different genres of the text is closely intervowen with the background knowledge of the content ofthe text.Elsewhere, Perfetti and Stafura (Perfetti, & Stafura, 2014) note that before teaching inferencing, understanding explicit information in the paragraph and connecting it to make meaning are taught. Moreover, some scholars point out that the central and elemental prerequisite in reading comprehension is vocabulary knowledge(Perfetti, Yang, & Schmalhofer, 2008;Yang, 2005, 2007). Grabe (Grabe, 2009) also adds that inferencing mostly depends on background and vocabulary knowledge among other attributes. Farr(Farr, 1992) concludes that comprehending reading abilities has evolved through Multiple-Choice (MC), cloze-test or open-ended questions in perceiving underlying abilities (Freedle, 1993; Nevo, 1989). However, Brown and Hudson (Brown, 2002), and Glaser (Glaser, 1994)emphasize that MC is a standard objective form of assessment particularly in Norm-Referenced Test (NRT). Because the manifestation of test takers' proficiency for making unbiased high-stakes decisions is possible on statistical assumptions. These findings collected in statistical analyses can receive their credits by fairness which could be possible through checking DIF.

**Differential item functioning**

DIF occurs when the probability of answering an item correctly is different across groups including age, cultural, gender differences, and so forth. To resolve this problem, DIF detection will ensure fairness to improve test validity in item performance (Kamata, 2004; Roussos, 2004). DIF exists "when an item's properties in one group are different from the item's properties in another group" (Furr, 2007). Thus, it is assumed that biased items function against the minority (focal) group such that they could not reveal their outperformance compared with the majority (reference) group. In the late 1980s, DIF was replaced with item biaswhich was associated with the unfair test items stemmed in social and political issues (Ellis, 2003).

In DIF studies, two types of DIF have been discussed in the literature including uniform and non-uniform DIF. The former refers to when an item is different across groups, whereas the latter

argues that non-uniform DIF occurs when an item differs across groups in terms of item difficulty parameters (Clauser, 1998).Basically, there exist two stages for DIF analysis. The first stage refers to statistical detection of DIF, and the second one identifies the sources of DIF and the underlying reasons why DIF has occurred. According to Van Nijlen and Janssen (Van Nijlen, 2011), identifying sources of DIF aid researchers' understanding of test construct.This perception is assumed as a necessary aspect of construct validation (Baghaei, 2016; Borsboom, 2004).

In CDA, DIF detection investigates the time when the probability of success on test items is different for examinees who are from different groups, but have the same attributes mastery profiles. This is a threat to the validity argument of the test. Technically, a test measures the samelatent trait for all test takers; it is also expected toentail the same difficulty level across different

populations.

To remove DIF problems, the Wald test needs to be run to analyze the status of each item in a high-stakes test. In the end, to address whether test takers' responses are a function of intended ability the following research questions are posited:

*$RQ_1$: Do reading comprehension test items of nationwide university entrance examination in PhD level flag DIF under CDA? If so, what is the effect size?*

*$RQ_2$: Does DIF in each test item affect the examinees' performance based on gender differences?*

**METHODS**

This study employed a sequential exploratory mixed method design to investigate a two-phase data collection at two different times.As it is common in sequential exploratory mixed methods design, the participants in quantitative study were larger, and they were not the same individuals who provided qualitative data.

**Participants**

The data for the present study were collected in two stages. The participants of the qualitative stage were PhD candidates of TEFL, andexperts in Applied Linguistics including 4 males and 9 females between the age ranges of 25 to 50.

As for the quantitative stage, 4200 test takers, who sought to pursue their studiesin Teaching English as a Foreign Language (TEFL), English Language Literature, English Translation, or Linguistics in PhD level at state universities, were randomly selected.They were both females and males including 61.3% females and 38.7 males. Participants had also completed MA degree in-English Translation, English Language Literature, English Language Teaching and Linguistics.

Generally, large sample sizes tend to present statistical differences meticulously with minor variations in sample's performance, which lead to reach solid findings and justifications.

**Instrumentations**

To develop a Q-matrix in qualitative phase of the present research, five reading comprehension attributes (Gao, 2010;Jang, 2009)were selected. The reading comprehension attributes included-vocabulary, syntax, extracting explicit information, connecting and synthesizing, and making inferences.

For the quantitative phase of the study, two reading comprehension passages of general English booklet were used to analyze ten test items. What needs to be considered was that the corpus of the study was provided by the National Organization for Educational Testing (NOET), which undertakes to administer high–stakes tests at both the undergraduate and graduate levels. The test understudy was administered by the NOET for the nationwide PhD admission test. And because of the critical role of high-stakes test results which necessitate strict confidentiality, a license was granted by theNOET to the researchers to receive the raw data.

**Procedure**

Each year, university entrance examination as a high-stakes test is scheduled to run in March. Test takers of the present study were required to answer 30 items including vocabulary and grammar (20 items), and reading comprehension (10 items) in 2016. The allotted time to complete the whole test was 45 minutes.

To generate information for diagnosis, and after a brief training session, participants in the qualitative phase of the study were invited to read reading comprehension passages, answer 10 test items, and verbalized their thoughts. Here, each student read each passage in a retrospective think-aloud session, and recounted the processes they used among the five provided attributes. To map test items onto particular attributes, it is critical to develop a Q-matrix. To construct a refined Q-matrix in think-aloud verbal protocol

analysis stage of the study, participants were asked to express their thoughts in a written open ended interview. Next, the panel of professors described earlier was invited to examine the extent to which each reading attribute resides in per test item. They were asked to rate how sure they were each attribute was necessary on a scale of one to five for each attribute.The results obtained used to construct a Q-matrix, in which there were items in a column and attributes appeared in a row.

Generally, it is assumed that the necessary

attributes to answer each test item show in 1, whereas the unnecessary attributes for each test item present in 0(Tatsuoka, 1983). (Table 1)

The constructed Q-matrix along with collected raw data in quantitative stage of the study was usedto measure the item and model fit indices. Then, the best fit model was selected. In the end, the Wald test was run through the application of GDINA rule to feed some arguments in R studio package of difR. In what follows, the results of data analyses were provided.

**RESULTS**

In the present study the R studio package of difR was utilized. Through the application of R studio, data processing was conducted meticulously.

To assess DIF, the first step was developing a Q-matrix. As for indicating desirable consensus among participants and experts' decisions, the

Kappa Coefficient of Agreement was estimated (k=0.78) which was approximately perfect. The

Kappa Coefficient of Agreement showed reliability of the Q-matrix. Then, the coded Q-matrix of participants and professors fed into the SPSS software by the researchers. Next, the relationship between coded attributes was checked through running Phi Correlation Coefficient of Agreement which there was negative relationship between attributes. As for validity of the Q-matrix, it was evaluated by experts' agreement to reach saturation.

In the next phase of analyses, by the application of R studio package, DINA and GDINA models were used to test model fit indices.

**Model and Item Fit Indices**

In CDA, selecting the best fit model has always been a challenge for practitioners. In the present study, attempts have been made to check model and item fit indices in advance. In this case, Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC),and -2log-likelihood (-2LL) were estimated. According to Lei and Li(Lei, 2016), AIC reports the best fit model in comparison to another model. In addition, R studio also shows the fit indices of the model in detail. Among them, $Mx^2$ has high power of detecting the fitness of selected model (Chen, & Thissen, 1997; Lei, 2016). In what follows,model and item fit indices were provided (Tables 2 & 3). Since this study first aimed at selecting

the best fit model in PhD nationwide admission test in 2016, AIC, BIC, -2LL, and $Mx^2$ were estimated based on GDINA and DINA models.

As displayed in Table 2, the lowest value was related to AIC in comparison to BIC. That is (AIC=32158) for selecting GDINA model. Moreover, the lowest value for selecting DINA model was obtained (AIC=32595). As a result, the structure of reading comprehension test items was more valid through selecting GDINA model comparing to DINA outputs.

To ensure, whether GDINA model fitted for the reading comprehension data, the results of $Mx^2$ determined a non-significant value of ($Mx^2$=4.23), p=1.00 for GDINA model, because they were almost closer to zero, whereas the values obtained for the DINA model were ($Mx^2$=142.18), p=1.00.

**Table 1.**

***Constructed Reading Comprehension Q-Matrix for each Test Item in 2016***

| Items | Vocabulary | Syntax | Extracting Explicit Information | Connecting and Synthesizing | Making Inferences |
|---|---|---|---|---|---|
| Q1 | 1 | 0 | 1 | 0 | 1 |
| Q2 | 1 | 0 | 0 | 1 | 1 |
| Q3 | 1 | 0 | 1 | 0 | 1 |
| Q4 | 1 | 0 | 1 | 0 | 0 |
| Q5 | 0 | 1 | 0 | 1 | 0 |
| Q6 | 1 | 1 | 0 | 1 | 0 |
| Q7 | 0 | 0 | 0 | 1 | 1 |
| Q8 | 1 | 0 | 1 | 1 | 1 |
| Q9 | 0 | 0 | 1 | 0 | 1 |
| Q10 | 1 | 0 | 1 | 0 | 1 |

**Table 2.**

***Model and Item Fit Indices based on GDINA Model in 2016***

| Reading Comprehension | AIC | BIC | -2Loglike | $Mx^2$ |
|---|---|---|---|---|
| GDINA Model | 32158 | 32817 | -15974.91 | 4.23 (p=1.0000000) |

**Table 3.**

***Model and Item Fit Indices based on DINA Model in 2016***

| Reading Comprehension | AIC | BIC | -2Loglike | $Mx^2$ |
|---|---|---|---|---|
| DINA Model | 32595 | 32924 | -16245.3 | 142.18 (p=1.0000000) |

The results obtained in comparing GDINA and DINA models revealed that GDINA model was the best fit model. It is expected that the saturated GDINA model would produce the best DIF result since this model is more parameterized. The following is the results which were investigated to measure the extent of DIF in reading comprehension section of PhD nationwide admission test items in 2016. (Table 3)

To estimate DIF, the Wald test was run to show whether a set of parameters was equal to some values. More specifically, de la Torre and Lee (de la Torre, & Lee, 2013) believe that the Wald test evaluate the fit of the model at the item level.Furthermore, the adjusted p-value was improved by means of the Bonferroni method to detect DIF items in CDM.

In Table 4, p-value indicated the typical significance level for the Wald statistic for items 1, 2, 3, 5, 6, 8, 9, and 10. The results of adjusted p-value through the Bonferroni showed that 2 items of the reading comprehension subtest did not have DIF under the fitted multiple group GDINA model. To address the first research question, based on the results of multiple groups GDINA model, 8 items had uniform DIF. However, the effect sizes measured for them were negligible or large.

**Table 4.**

***DIF Detection through the Wald Statistic in 2016***

| Items | Wald Statistic | df | P-Value | Adjusted P-Value |
|---|---|---|---|---|
| 1 | 119.60 | 8 | 0.00 | 0.00** |
| 2 | 22.07 | 8 | 0.00 | 0.04*** |
| 3 | 28.40 | 8 | 0.00 | 0.00** |
| 4 | 5.82 | 4 | 0.21 | 1.00* |
| 5 | 43.17 | 4 | 0.00 | 0.00** |
| 6 | 90.74 | 8 | 0.00 | 0.00** |
| 7 | 11.87 | 4 | 0.01 | 0.18* |
| 8 | 90.93 | 16 | 0.00 | 0.00** |
| 9 | 44.85 | 4 | 0.00 | 0.00** |
| 10 | 72.69 | 8 | 0.00 | 0.00** |

*Note:* adjusted p-values are based on the Bonferroni correction.

*Note:* Effect Size Evaluation is based on

* Non-significant

** Large***Negligible

**Table 5.**

***Skill Mastery Probabilities Based on Gender Difference***

| Skills<br>Genders | Vocabulary | Syntax | Extracting Explicit Information | Connecting and Synthesizing | Making Inferences |
|---|---|---|---|---|---|
| Females | 0.27 | 0.25 | 0.24 | 0.40 | 0.30 |
| Males | 0.33 | 0.25 | 0.23 | 0.47 | 0.36 |
| Difference | -0.05 | -0.00 | 0.00 | -0.07 | -0.05 |

To answer the second research question, item and attribute parameters were estimated by the application of GDINA model.Here, male and female different performances determined based on class probabilities. The results of the differences presented the higher or lower chance of mastery or mastery probabilities in reading comprehension attributes. Table 5 presents males outperformance in comparison to females. That is to say, except for extracting explicit information, in all other attributes males could performed higher than females.

**DISCUSSION**

This research illustrated the application of the CDA in language assessment and,more specifically, in the analysis of DIF in reading comprehension section of a PhD national admission test. That is to say, construct validity as the primary and crucial factor in diagnostic assessment is taken into account, because validity is assumed in terms of accuracy of measurement. The validity of a high-stakes test is estimated to find whether test items suspected DIF. In addition, gender differences are measured to determine skill mastery probabilities of test takers. In practice, to answer the first research question, model and item fit indices showed that the GDINA model fitted the data well, whereas DINA model did not fit. To ensure correct identification of attributes to the items, the acceptable fit of the data to CDA wasacceptable evidence for available findings(Chen, de la Torre, & Zhang, 2013). Regarding the second research question, the results of DIF showed that some items flagged DIF in favor of males. To identify the quantitative differences in each test item in terms of DIF, adjusted p-value were measured. It is possible to confirm that 8 out of 10 items (including items 1, 2, 3, 5, 6, 8, 9, and 10) suspected a large or negligible DIF.These results reveal that females are weaker in the present study, which is not in line with previous study on gender differences in EFL Iranian test takers (Farashaiyan, 2012). The reason that females received lower probabilities of mastery was that the passage provided for the nationwide university admission test argued women underrepresentation. According to Khodaii (Khodaii, 2009), test designers are required to develop materials which do not convey any bias for test takers in test venue; thus, all should have the same chance to elicit response. However, in this year, DIF suspected against female. Thus,there are dependencies among the attributes of reading comprehension which result in designing unfair test items.

As a result, it is vital to promote CDA further in ELT program. In this way, statistical feedback provided for stakeholders with different mastery profiles in reading comprehension skills is critical. Howeverproviding diagnostic feedbackfor stakeholders demands considering test use in social systems. Behuniak (Behuniak, 2002) and Shohamy(Shohamy, 2001)emphasize "consumer-reference testing" and "use-oriented testing" in language testing and teaching.

Stiggins and Conklin (Stiggins, 1992) emphasize that typical teachers spend between one-third and one-half of their class time on assessment activities, in spite of the fact that teachers' knowledge about assessment matters have been limited (Christie, 1991; Louden, 2005; Matters, 2006). To remove such a problem, through practicing cognitive diagnosis the problem in assessment could be lessened. In this way, diagnosis, treatment, and assessment can be assumed as elements of a loop in which facilitators could diagnose problems of students, engage them in problem solving activities, and assess their learning in multiple tasks. In such a way, teachers can also focus on what students lack knowledge of it. In addition, CDA assists teacher training courses. That is to say, teachers in training courses can become familiar with some diagnostic information about skill mastery or non-mastery of test takers. This information can aid teachers to find how they can successfully encounter with students problems and facilitates process of learning in real situation. As a result, language teachers should be trained such that they can perceive how to deal with some circumstances either when test

takers in different groups do not have equal opportunity to learn the materials being tested, or even when a fair test can be used and interpreted unfairly.

Finally, to develop materials fairly, assessment specialists and materials developers need to collaborate and to design materials which convey diagnostic specifications. In practice, this may result in providing some kinds of activities to facilitate skill development, and improve expected outcomes of learning. This can be identified as curriculum goals. Because, it would necessitate considering test takers' performances along with features of context which are appropriate for future tests(Wilson, 2012). Furthermore, Hughes(Hughes, 1989), and Lee and Sawaki (Lee, 2009a) assert that this is the responsibility of program developers and administers to determine the type of materials that teachers can clearly use to perceive students strengths and weaknesses. It is also possible through collecting separate diagnostic information for different proficiency levels at classroom level and high-stakes program level. This may assist in the development of some lesson plans revealed students different needs. Therefore, policy makers could take into consideration the accountability purposes aiming to apply in assessment for learning (Black, & William, 2004; Black, Harrison, Lee, Marshall, & William, 2003; Council, 2001;

Harlen, 2005; Kellis, 2002). In sum, the conduction of this study was not without limitations. One of the most sophisticated issues in the present study is the sociopolitical factors which affect test development and use of high-stakes tests. However, due to practical considerations these issues leave untouched. It is also unclear whether this differential performance of test takers is due to different abilities of examinees or artifacts of the assessment instrument itself.

## CONCLUSIONS

The goal of present research isto show skill mastery profiles of stake holders in a real educational context. This could provide useful information about test takers strengths and weaknesses in reading abilities unless test items suspected to DIF. The present study presented that CDA could identify both unfair test items, and underlying performances of both female and male in a high-stakes test.

It would be necessary to emphasize some theoretical and practical dilemmas which left some future research areas open to further investigation. This study can be extended to using different attributes, and engaging more participants of other majors. In this way, information which would be gathered from various test takers in think-aloud verbal protocols could be used more authentically for developing a Q-matrix. This may result in developing fair high-stakes English language tests under CDA. This study also underwent reversed engineering (retrofitting) analysis in CDA. In addition, there is as yet no standardized method for Q-matrix development based on the extracted attributes from related literature, experts' judgments, and students' think-aloud verbal protocol analyses. Selecting the best fit model other than the ones applied in the present study is suggested. Construct under- representation and construct irrelevant variables are also vital to be observed in the process of test development. In the end, future research is needed to design a practical, useful, and in effect PhD nationwide university admission tests in Iran which would conform to the standards of fairness.

**References**


Alderson, J. C. (1990a). Testing reading comprehension skills (Part one). Reading in a Foreign Language, 6(2), 425–438.

Alderson, J. C. (1990b). Testing reading comprehension skills (Part two): Getting students to talk about taking a reading test (A pilot study). Reading in a Foreign Language, 7(1), 465–503.

Alderson, J. C. (2000). Assessing reading. Cambridge: Cambridge University Press.

Alderson, J. C. (2005). Diagnosing foreign language proficiency: The interface between learning and assessment. London: Continuum.

American Educational Research Association (AERA), A. P. A. A., & National Council on Measurement in Education (NCME). (1999). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

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

Anderson, R. C., & Pearson, P. D. (1988). A schema-theoretic view of basic processes in reading comprehension. In J. D. D. E. E. E. In P. L. Carrell (Ed.), Interactive approaches to second language reading (pp. 37-55): Cambridge: Cambridge University Press.

Aryadoust, V. (2011). Application of the fusion model to while-listening performance tests. SHIKEN: JALT Testing and Evaluation SIG Newsletter, 15(2), 2-9.

Baghaei, P., & Ravand, H. (2015). A cognitive processing model of reading comprehension in English as a foreign language using the linear logistic test model. Learning and Individual Differences, 43, 100-105.

Baghaei, P. T.-Y., M. (2016). The logic of latent varaible analysis as validity evidence in psychological measurement. The Open Psychology Journal, 9, 168-175.

Bechtoldt, H. P. (1951). Selection. In I. S. S. Stevens (Ed.), Handbook of experimental
psychology (pp. 1237–1267). New York:
Wiley.

Behuniak, P. (2002). Consumer-referenced testing. A Phi Delta Kappan, 199-207.

Black, P., & William, D. . (2004). The formative purpose: Assessment must first promote learning. In I. M. Wilson (Ed.), Towards coherence between classroom assessment and accountability. Chicago, IL: National Society for the Study of Education.

Black, P., Harrison, C., Lee, C., Marshall, B., & William, D. (2003). Assessment for learning: Putting it into practice. Buckingham: Open University Press.

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2004). The concept of validity. Psychological Review, 111, 1061-1071.

Brown, J. D. H., T. (2002). Criterion- referenced language testing. Cambridge: Cambridge University Press.

Carrell, P. L. (1988). SLA and classroom instruction: Reading. Annual Review of Applied Linguistics, 9, 223–242.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cogniitve diagnostic modeling. Journal of Educational Measurement, 50, 123-140.

Chen, H., & Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. Language Assessment Quarterly, 13(3), 218-230.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. Journal of Educational and Behavioral Statistics, 22(3), 265–289.

Chomsky, N. (1957). Syntactic structures. The Hague: Mourton & Company.

Christie, F., Devlin, B., Freebody, P., Luke, A., Martin, J. R., & Threadgold, T. (1991). Teaching English literacy: A project of

national significance on the preservice preparation of teachers for teaching English literacy (Vol. 1): Canberra: Department of Employment, Education and Training.

Clapham, C. M. (1996). The development of IELTS: A study of the effect of background knowledge on reading comprehension. Cambridge: Cambridge University Press.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. Educational Measurement: Issues and Practice, 17(1), 31-44.

Cooper, M. (1984). Linguistic competence of practiced and unpracticed non-native readers of English. In I. J. C. A. A. H. Urquhart (Ed.), Reading in a foreign language (pp. 122-138). London: Longman.

Council, N. R. (2001). Knowing what students know: The science and design of educational assessment. Washington: National Academy Press.

Cronbach, L. M., P. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281–302.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. Psychometrika, 69(3), 333-353.

de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. Journal of Educational Measurement, 50(4), 355-373.

Ellis, B. B., & Raju, N. S. (2003). Test and Item Bias: What they are, What they aren't, and How to detect them. Educational Resources Information Center (ERIC).

Farashaiyan, A. T., K. H. (2012). On the relationship between pragmatic knowledge and langugae proficiency among Iranian male and female undrgraduate EFL learners. 3L: Language, Linguistics, Literature. The Southeast Asian Journal of English Langauge Studies, 18, 33-46.

Farr, R. (1992). Putting it all together: Solving the reading assessment puzzle. The Reading Researcher, 46(1), 26–37.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: implications for construct validity. Language Testing, 10(2), 133-170.

Furr, M. R., & Bacharach, V. R. (2007). An introduction: Psychometrics: Thousand Oaks, CA: SAGE.

Gao, L., & Rogers, W. T. (2010). Use of tree-based regression in the analyses of L2 reading test items. Language Testing, 28(2), 1-28.

Gaylord, R. H. (1955). Conceptual consistency and criterion equivalence: a dual approach to criterion analysis. Unpublished Manuscript as cited in Cronbach & Meheel (PRB Research Note No. 17). Copies obtainable from ASTIA-DSC, AD-21 440.

Glaser, R. (1994). Instructional technology and the measurement of learning outcomes: Some questions. Educational Measurement: Issues & Practice, 13(4), 6-8.

Goodman, K. S. (1967). Reading: A psycholinguistic guessing game. Journal of the Reading Specialist, 6(1), 126-135.

Gough, P. B. (1972). One second of reading. In I. J. F. K. I. G. Mattingly (Ed.), Language by ear and by eye. Cambridge, Mass.: MIT Press.

Grabe, W. (1988). Reassessing the term ‘interactive'. In J. D. In P. L. Carrell, & D.E. Eskey (Ed.), Interactive approaches to second language reading (pp. 57-69). Cambridge: Cambridge University Press.

Grabe, W. (1991). Current developments in second language reading research. TESOL Quarterly, 25(3), 375–406.

Grabe, W. (2000). Developments in reading research and their implications for computer-adaptive tests of reading. Cambridge: Cambridge University Press.

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

York, NY: Cambridge University Press.

Harlen, W. (2005). Teachers' summative practices and assessment for learning-tensions and synergies. The Curriculum Journal of Educational and Behavioral Statistics, 16(2), 207-223.

Hemmati, S. J., Baghaei, P., & Bemani, M. (2016). Cognitive Diagnostic Modeling of L2 Reading Comprehension Ability: Providing Feedback on the Reading Performance of Iranian Candidates for the University Entrance Examination. International Journal Langauge Testing, 6(2).

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. Psychometrika, 74(2), 191-210.

Hughes, A. (1989). Testing for language teachers. Cambridge: Cambridge University Press.

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)].

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.

Kamata, A., & Vaughn, B. K. (2004). An introduction to differential item functioning analysis. Learning Disabilities: A Contemporary Journal, 2(2), 49-69.

Kellis, M., & Silvernail, D. (2002). Considering the place of teacher judgment in Maine's local assessment systems: Maine: Center for Educational Policy, Applied Research, and Evaluation, University of Southern Maine.

Khodaii, E. (2009). Effective Factors in Passing MA Entrance Examination. Higher Education Studies and Planning Quarterly, 54, 15-34.

Kim, A.-Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. Langauge Testing, 322(2), 227-258.

Kintsch, W. v. D., T. A. (1978). Toward a model of text comprehension and production. Psychological Review, 85(34), 363–394.

LaBerge, D. S., S. J. (1974). Toward a theory of automatic information processing in reading. Cognitive Psychology, 6(2), 293-323.

Langer, J. A. A., R. L. (1992). Curriculum research in writing and reading. In I. P. Jackson (Ed.), The handbook of curriculum research (pp. 687–725). New York: Macmillan.

Lee, Y. W., & Sawaki. (2009a). Cognitive diagnosis and Q-matrices in language assessment. Language Assessment Quarterly, 6(1), 169-171.

Lei, P. W., & Li, H. (2016). Fit indices' performance in choosing cognitive diagnostic models and Q-matrices.Applied Psychological Measurement. 40(6), 405-417.

Li, H. (2011). Evaluating langauge group differences in the subskills of reading using a cognitive diagnostic modeling and differential skill functioning approach. Unpublished doctoral disserttaion, Penn State University, State College, PA.

Louden, W., Rohl, M., Gore, J., McIntosh, A., Greaves, D., Wright, R. (2005). Prepared to teach: An investigation into the preparation of teachers to teach literacy and numeracy. Canberra: Department of Education, Training and Youth Affairs.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. Language Testing, 10(3), 211-234.

Matters, G. (2006). Assessment approaches in Queensland senior science syllabuses. A

report to the Queensland Studies Authority. Brisbane: ACER.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. Educational Measurement. Issues and Practice, 14(4), 5-8.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. Language Learning, 6(2), 199-215.

Perfetti, C. A., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. Scientific Studies of Reading, 18(1), 22-37.

Perfetti, C. A., Yang, C-L., & Schmalhofer, F. (2008). Comprehension skill and word-to-text processes. Applied Cognitive Psychology, 22(3), 303-318.

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.

Ravand, H. (2015). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. Journal of Psychoeducational Assessment, 1-8.

Ravand, H. (2016). Application of a cogniitve diagnostic model to a high stakes reading comprehension test. Journal of Psychoeducational Assessment, 34(8), 782-799.

Ravand, H., & Robitzsch, A. . (2015). Cognitive diagnostic modeling using R. Practical Assessment, Research and Evaluation, 20(11), 1-12.

Ravand, H., Barati, H., & Widhiarso, W. (2012). Exploirng diagnostic capacity of a high-stakes reading comprehension test: A pedagogical demonstration. Iranian Journal of Language Testing, 3(1), 12-37.

Rost, D. H. (1993). Assessing different components of reading comprehension: Fact or fiction. Language Testing, 10(1), 79-92.

Roussos, L., & Stout, W. (2004). Differential item functioning analysis: Detecting DIF items and testing DIF hypotheses. In I. D. Kaplan (Ed.), The value Sage handbook for social sciences (pp. 107-115): Newbury Park, CA: Sage.

Rumelhart, D. (1977). Toward an interactive model of reading. In I. S. Domic (Ed.): Attention and performance (VI). Hillsdale, N.J.: Erlbaum.

Rumelhart, D. (1980). Schemata: the building blocks of language. In B. C. B. In R. J. Spiro, & W. F. Brewer (Ed.), Theoretical issues on reading comprehension (pp. 33-58): Hillsdale, New Jersey: Erlbaum.

Salager-Meyer, F. (1991). Reading expository prose at the post-secondary level: The influence of textual variables on L2 reading comprehension (a genre-based approach). Reading in a Foreign Language, 8(1), 645–662.

Shohamy, E. (2001). The power of tests: A critical perspective on the use of language tests. Harlow, England: Longman.

Smith, F. (1971). Understanding reading. New York: Holt, Rinehart and Winston.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In I. R. L. Linn (Ed.), Educational measurement (pp. 263-331). New York: American Council on Education/Macmillan.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. Reading Research Quarterly, 16(1), 32-71.

Stiggins, R. J., & Conklin, N. F. (1992). In teachers' hands: Investigating the practices of classroom assessment. Albany, practices NY: State University of New York Press.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement: Issues & Practice, 20(4), 345-354.

Tatsuoka, K. K. (1990). Toward an integration of item–response theory and cognitive error diagnosis. . In R. G. N. Frederiksen, A. Lesgold, & M. G. Shafto (Eds.) (Ed.),

Diagnostic monitoring of skill and knowledge acquisition (pp. 453-488). Hillsdale, NJ, US: Lawrence Erlbaum Association, Inc.

Techniques, T. R. f. P. T. a. D. (1954). Psychological Bulletin, Supplement. Psychological Bulletin, Supplement, 51(2), 1-38.

Templine, J. L., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. Psychometrika, 79, 317-339.

Van Nijlen, D., & Janssen, R. . (2011). Measuring mastry across grades: An application to spelling validity. Applied Measurement in Education, 24, 367-387.

Watson, J. (1913). Psychology as the behaviorist views it. Psychological Review, 117-158.

Wilson, M. R., Bejar, I., Scalise, K., Templin, J., Wiliam, D., & Irribarra, D. T. (2012). Perspectives on methodological issues. In B. M. In P. Griffin, & E. Care (Ed.), Assessment and teaching of 21st century skills (pp. 67-141). Dordrecht, the Netherlands: Springer.

Yang, C.-L., Perfetti, C. A., & Schmalhofer, F.(2005). Less skilled comprehenders' ERPs show sluggish word-to-textintegration processes. Written Learning & Literacy, 8(2), 233-257.

Yang, C.-L., Perfetti, C. A., & Schmalhofer, F.(2007). Event-related potential indicators of text integration across sentence boundraries. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(1), 55-89.

**Biodata**

**Ms Niloufar Shahmirzadi** is a Ph.D. candidate of Applied Linguistics from Islamic Azad University at Central Tehran. She has published some articles and books, and attended some national and international conferences. She is a member of the Young Researchers and Elite Club. Currently she is a lecturer. Her major areas of research lie in Language Testing and Assessment, as well as Performance Assessment. She has published in *ILTA* Newsletter 2019.
Email: niloufar_shahmirzadi83@yahoo.com

**Dr Masood Siyyari** is an Assistant Professor at Science and Research Branch, Islamic Azad University, Tehran. His main areas of interests are Quantitative Data Analysis in Applied Linguistics, Language Testing and Assessment, as well as SLA. He has published and presented in both national and international journals and conferences.
Email: m.siyyari@srbiau.ac.ir

**Dr Hamid Marashi** is anAssociate Professor of Applied Linguistics, Islamic Azad University at Central Tehran and Editor-in-Chief of the Journal of Language and Translation. He currently teaches graduate and postgraduate courses with his main areas of research interest including innovative teaching practices and learner variables. He has published in international academic journals (including TESOL Journal and Language Learning Journal) and also presented in international conferences.
Email: hamid.marashi@iauctb.ac.ir

**Dr Masoud Geramipour** is an Assistant Professor in the Department of Curriculum Studies, Kharazmi University. He has taught research methodology and psychometrics courses to educational research students since 2010.
Email: mgramipour@khu.ac.ir