Language Testing in Focus: An International Journal

Language Testing in Focus
An International Journal
LTiF
ISSN: 2717-9087

www.EUROKD.com

2023 (8)

Cognitive diagnostic assessment of reading comprehension for high-stakes tests: Using GDINA model

Niloufar Shahmirzadi*, Hamid Marashi
Department of Foreign Languages, Tehran Central Branch, Islamic Azad University, Tehran, Iran

ABSTRACT

Cognitive diagnostic assessment (CDA) is used to study cognitive and educational psychology, and designed to diagnose the underlying abilities of test takers in comprehension language skills such as reading comprehension. Through applying CDA, a test has undergone accurate studies to remove biased test items which yield great impact on individuals, educational systems and societies. In this case, psychometric statistical analyses were applied, Differential Attribute Functioning (DAF)) was also used to detect the probability of the mastery of attributes among test takers, and Differential Item Functioning (DIF) was estimated to show item performance among different candidates in terms of gender, their GPAs in BA, and MA degrees. The randomly selected participants of this study were 7,420 females and males sitting for the nationwide PhD admission test to pursue their education in Applied Linguistics. Moreover, a Q-matrix was developed, data were fed into R studio software, and the Generalized Deterministic Inputs, Noisy "and" Gate (GDINA) model was run. The results of the study flagged large DIF in gender group in 2019; and in gender, BA, and MA groups in 2020. In sum, this study is an attempt to raise the awareness of test developers to shed light on the critical discursive sources of inequity and bias. The implication of this study can provide pedagogically useful diagnostic information for test designers and teachers since a proficiency test needs to be valid, reliable, and fair in the context of high-stakes tests so that it would lead to positive changes.

Keywords
Cognitive Diagnostic Assessment, Differential Attribute Functioning, Differential Item Functioning, Reading Comprehension

Received
16 June 2023
Received in revised form
03 September 2023
Accepted
16 September 2023

Correspondence concerning this article should be addressed to:
niloufar_shahmirzadi83@yahoo.com

Introduction

In traditional assessments which originated in Item Response Theory (IRT) or Classical Test Theory (CTT), test takers' scores are determined by a single latent proficiency continuum to

https://doi.org/10.32038/ltf.2023.08.01

compare ability, identify level of proficiency, differentiate passing or non-passing, locate the difficulty of items, select a program, and depict between rather than within item multi-dimensionality regardless of revealing diagnostic information. The common unidimensional IRT models apply any skills to report on a continuum which can neither reveal strengths or weaknesses of test takers, nor permit to obtain a profile of skills mastery. Therefore, there is a need to measure latent attributes or sub-skills in a fine-grained size to indicate which specific skills have or have not been mastered (de la Torre, 2009). Alderson (2005) accentuates that the finer the grain size, the more detailed the information might delineate. Within recent decades, cognitive diagnostic models (CDMs) as typically used *for* learning as opposed to *of* learning (Jang, 2008) aiming to classify individuals based on their item response patterns. That is, it can provide diagnostic information and instruction. According to Rupp and Templin (2008), "CDMs is useful for modeling observable categorical response variables and contains unobservable or latent categorical predictor variables" (p. 226).

Brown (2004) believes that in standardized testing, a comprehensive definition of competency depicts underlying language abilities of test takers. Interestingly, these advances in language assessment and cognition require a direct novel approach to diagnostic assessment (Kunnan & Jang, 2009; Jang, 2009) which attempts to make a connection between the skill competencies of test takers in certain attributes and characteristics of test items in order to elicit certain skills and to develop a Q-matrix. Q-matrix shows the relationship between a test item and its required skills. In Q-matrix construction, the term skills also known as attributes, knowledge, abilities, processes, strategies, and sub-skills which are used interchangeably. In detail, attributes refer to "any latent knowledge or abilities to complete a task" (Buck & Tatsuoka, 1998, p. 121). As a result, this would lead to multiple attributes within an item. Furthermore, because attributes are limited to the number of restricted latent class in a test item, they are also called restricted latent class (Haagenars & McCutcheon, 2002).

Loaded attributes in each test item are coded to develop a Q-matrix (Tatsuoka, 1983, 1990). Then, each response to a test item with a Q-matrix can be analyzed and translated to ease the score report. However, there are still some conceptual difficulties to translate these attributes into a meaningful score for cognitive diagnostic assessment (CDA). Stiggins (2002) and Pellegrino, et al. (1999, p. 335) also note that assessment should determine successful learning rather than the status of learning to optimize "interpretative, diagnostic, highly informative, and potentially prescriptive" learning. Thus, recent endeavors have been made to incorporate cognitive structures in psychometric models and assessment to develop CDMs. This can assist in analyzing learning in an actively richer psychometric context (Rupp & Templin, 2008; Rupp, et al., 2010) because CDMs show test takers' abilities in different cognitive domains and then provide diagnostic feedback.

Most of the previous studies focused on a single predetermined model in CDMs with reading comprehension (Yi, 2012). In current practices, however, one major difficulty is the selection of a model from among a large number of CDMs (Jiao, 2009). Generally speaking, CDMs are classified into specific and general models, for example; (DINO, NIDO), (DINA, NIDA), (ACDM, C-RUM, LLM), (GDM), (LCDM), and (GDINA) models. General models of various formulations have been proposed to allow for both types of relationship within a model such as GDM (Von Davier, 2005), LCDM (Henson, Templin, & Willse, 2009), and GDINA (de la Torre, 2011). Specifically, Generalized DINA named as GDINA is a relax model that can measure all the required attributes for response to an item. That is to say, when none of the required attributes has been mastered, it can estimate the probability of correct response to an item unlike DINA model. GDINA model also depicts different relationship between skills including all main and interaction effects. The other crucial decision in choosing a model is that of adopting compensatory (non-conjunctive) or non-compensatory (conjunctive) model (Roussos, Templin, & Henson, 2007). To name a few other models, DINO (Templin & Henson, 2006), and C-RUM (Hartz, 2002) are applied for compensatory models, and DINA (Junker & Sijtsma, 2001), and NC-RUM (DiBello, Stout, & Roussos, 1995; Hartz, 2002) are used for non-compensatory models. However, de la Torre (2011, p. 179) believes that "whether different CDM formulations represent different classes of models, or to what extent these models are relevant to one another" have not entirely been crystalized.

Interestingly, English language skills researchers hold the view that reading comprehension is a compensatory interactive skill to the extent that non-native speakers can take advantage of their first language ability and second language knowledge to compensate for deficiencies in one or more attributes (Bernhardt, 2005; Goldsmith-Phillips, 1989; Stanovich, 1980). Henceforth, compensatory models in CDM may move toward more straightforward and meaningful interpretation because they are flexible enough to show the relationship among all attributes, and to analyze fair designed items within multiple groups (Johnson, Lee, Sachdeva, Zhang, Waldman, & Park, 2013). Moreover, in this process of item analysis, DIF detection explains as the matter of item invariance. Zumbo (2007) defines DIF in involving item bias in high-stakes questions and invariance in focal group (minority) and reference group (majority). Here, item bias is replaced by differential item functioning. By this, the main concern is flagging DIF statistically and defining true difference in test takers' abilities where they emerged in performance. The main effects of group difference and the interaction effect of group by ability (uniform and non-uniform DIF, respectively) are also examined in DIF detection. It is worth noting that the most common statistical method for DIF is estimated by signed area tests (mainly on uniform DIF) and unsigned area (non-uniform DIF), and nested model testing via a likelihood ratio test. Later, multidimensional DIF detection reflects beyond underlying interest in a test (Ackerman, 1992). In detail, some drawbacks including inequity in testing irrespective of a possible threat to internal validity (Zumbo, 2007) may result in attempts to use of DIF. In response, developing a generalized way of DIF measurement would be possible. Accordingly, the extent to which DIF leads to biased attributes can also be scrutinized by differential attribute functioning (DAF) to show equating mastery on attributes across groups. In this case, some previous studies conducted to validate mostly high-stakes tests (e.g., Hemati & Baghaei, 2020; Ketabi, et al., 2021; Roohani Tonekaboni, et al., 2021; Shahmirzadi, et al., 2020 a, b; Shahmirzadi, 2023; Tabatabaee-Yazdi, et al., 2021); however, there is a need to study validity of test items with regard to measuring DIF and DAF. Therefore, this research aimed to measure DIF within items and check for DAF in terms of gender, BA, and MA in two years. In so doing, if some degree of DIF was found in assessment, item modification would be required prior to administration. All the aforesaid issues are examined in terms of this PhD national admission test. To fulfill the above-mentioned objective, the following research questions were raised:

RQ$_1$: Do PhD nationwide admission tests flag DAF and DIF in each test item in terms of gender, BA and MA degrees' GPA? If so, what is the effect size?

RQ$_2$: What is the contribution of this study to the test under investigation per se?

Methodology

Participants in the Qualitative Phase

Five PhD candidates majoring in Applied Linguistics were recruited by the researchers to participate in a think-aloud verbal protocol analysis. The participants were two males and three females with the average age of 35 studying at Islamic Azad University. After a brief training session to train how to code item attribute relationships among the attributes which were developed by Gao and Rogers (2010), and Jang (2009), each student identified 5 attributes in approximately 20 minutes for 10 reading comprehension questions. The reading comprehension attributes included vocabulary, syntax, extracting explicit information, connecting and synthesizing, and making inferences. Following this session, a follow-up open-ended structured written interview based on critical thinking dispositions (adapted from Hughes & Jones, 1988) was conducted so that the participants could clarify their statements on each item or attribute critically. Then, a panel of eight professors of Applied Linguistics comprising two males and six females who enjoyed 20 to 30 years of teaching experience at Islamic Azad University was invited to examine the extent to which each reading attribute resides in each test item. Having reviewed the reading questions, the professors immediately verbalized their thoughts in an open ended structured written interview with regards to the importance of using attributes in each test item. Ultimately, a refined coded scheme was used to develop a Q-matrix considering the extant literature.

Here, interactions between "cognitive skills and test items" are presented in a Q-matrix (Jang, 2009, p. 214). Thenceforth, de la Torre (2009, p. 2) asserted that, "it is a cognitive design matrix that explicitly identifies the cognitive specification for each item". In Q-matrix development 1s indicate that the attributes are required for the item, and 0s indicate that the attributes are not required. Here, eliciting sufficient skills is vital for statistical analysis (DiBello, et al., 1995; Jang, 2005, 2009; Yang & Embretson, 2007) to construct a solid Q-matrix. Thus, through running the Phi Correlation Coefficient of Agreement, correlation between attributes was checked.

The results revealed that there was a correlation between vocabulary and extracting explicit information, vocabulary and making inferences, syntax and extracting explicit information, syntax and making inferences, extracting explicit information and connecting and synthesizing, and connecting and synthesizing and making inferences in test 2019. In general, there were 4/15 (26.66%) strong correlation coefficients, 5/15 (33.33%) moderate correlation coefficients, and 6/15 (40%) weak correlations. In test 2020, there was a correlation between vocabulary and syntax, syntax and connecting and synthesizing, extracting explicit information and connecting and synthesizing; and there was a correlation between vocabulary and extracting explicit information, vocabulary and connecting and synthesizing, vocabulary and making inferences, syntax and extracting explicit information, syntax and connecting and synthesizing, and syntax and making inferences, respectively. There were 1/15 (6.66%) strong correlation coefficients, 1/15 (6.66%) moderate correlation coefficients, and 13/15 (0.86%) weak correlations in 2020.

To delve into the process of Q-matrix development, there is disagreement in rating attributes among raters and participants since an expert's ability to judge is above that of participants (Leighton & Gierl, 2007). To ensure a desirable consensus among participants and experts' decisions, the Kappa Coefficient of Agreement was estimated. The coded Q-matrix of participants and experts in each year was fed into the SPSS software separately in order to resolve this disagreement. Then, the command was run to observe inter-rater reliability. There was substantial agreement between the two estimations, k = .78 in 2019. And, there was also almost perfect agreement between the two diagnoses, k = 1.00 in 2020.

Besides, usually the reading passage tests were mostly long with less frequent vocabularies, and complicated sentences which made the process of Q-matrix development difficult. This may demand an accurate estimate of the average correlation of all items that pertain to a certain construct by running Cronbach's alpha to ascertain the reliability of a psychometric test. In this phase, the results displayed that there was almost a good reliability, α=.52 in 2019. However, the items were not reliable enough since there was low reliability if the items were deleted. In 2020 there was an acceptable reliability, α=.75. Having run statistical analyses, the researchers had to particularize the required attributes in a Q-matrix numerically (Tatsuoka, 1983, 1990). To do so, the finalized Q-matrices with five attributes in responding to each item were developed in two consecutive years as follows (see Tables 1 and 2).

Table 1
Developed Reading Comprehension Q-Matrix for per Test Item in 2019

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

Developed Reading Comprehension Q-Matrix for per Test Item in 2020

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

Participants in the Quantitative Phase

For quantitative phase, data were collected from the nationwide PhD admission test administered by the National Organization for Educational Testing (NOET). A total of 7,420 participants were randomly selected from two consecutive years including 2019 and 2020. The sample consisted of both females and males with the average age of 37.5. All participants had applied for a graduate degree in Applied Linguistics.

The General English test, which was adopted in this study, was due to be administered annually in March. It comprised of two subsections including vocabulary and grammar (20 items), and reading comprehension (10 items). The participants were supposed to complete the General English section in 45 minutes. Specifically, the reading comprehension module was based on two academic texts followed by 10 multiple choice items which were scored dichotomously. It is worth mentioning that the collected data were strictly treated confidential, and a license was granted by the NOET to the researchers.

Design of the Study

To meet the goal of this research, qualitative and quantitative data were used. A sequential exploratory mixed method design was also applied to investigate a two-phase data collection at two different times. Regarding the data analyses, data were fed into SPSS and R Studio software.

Results

Model Fit

As with any statistical model to check the best model fit selection, it is necessary to estimate the model fit indices (Brown & Hudson, 2002), and the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarzer, 1976) as the two main criteria. In these two criteria, the smallest value for the relative fit index is preferred (Li, et al., 2015; Rupp, et al., 2010). Akaike (1974) proposed that when comparing various models, the best-fitting model was the one with the lowest AIC and BIC values.

In 2019, the log–likelihood=-15960.88 led to the best-fitting model in the BA group, because this year presented the lowest value compared to the other years. Comparing AIC and BIC values among gender, BA and MA classifications, the lowest AIC=32130 and BIC=32789 values along with the best-fitting model for GDINA were obtained for the BA group. Thus, the structure of the reading comprehension test items was more valid in this group than in the other two groups. The log–likelihood=-13848.11 in 2020 also led to the best-fitting model in the BA group, because of its lowest value compared to another year. Comparing AIC and BIC values among gender, BA, and MA classifications, the lowest AIC=27908 and BIC=28552 values along with the best-fitting model for GDINA were obtained for the BA group. Thus, the structure of the reading comprehension test items in the BA group was more valid than in the gender and MA groups.

Item Fit Statistics

To gain meaningful results in any statistical model, it is required to check the fit of the item to be meaningful. This can be ascertained in two ways including checking the fit of the model with the data (i.e., absolute fit), and comparing the model with the other models (i.e., relative fit). According to Li et al. (2015), the absolute model fit measures both the absolute model fit and model predicted values. Some of its recommended techniques are Mean Absolute Difference (MADaQ3), Root Mean Square Error of Approximation (RMSEA), Chi square (Mx^2), Mean Absolute Difference for the Item-Pair Correlations (MADcor), and MADQ3 statistic.

Chen and Thissen (1997) believed that Mx^2 is the test global model fit which is the mean difference between the model predicted and the observed response frequencies. If CDM fits the data properly, the Mx^2 is expected to be 0 within each latent class "to predict the observed response pattern" (Rupp et al., 2010, p. 269). MADcor statistic was the observed and the model-predicted item correlations (DiBello, et al., 2007). And, the RMSEA is used for the item parameters. It showed fitness of items with the adopted model that is GDINA. Here, there were three ranges from 0 to 1 consisting of RMSEA<0.05 (good fitness of item with model), RMSEA>0.1 (weak fitness of item with model), and 0.05<RMSEA<0.1 (medium fitness of item with model). RMSEA estimated the difference between examined and the hypothetical model where every component in the model is related to every other component, and it is based on the covariance matrices.

In 2019 the (*RMSEA: Gender=0.012, BA=0.01, MA=0.012*), and in 2020 (*RMSEA: Gender=0.016, BA=0.014, MA=0.015*) were RMSEA<0.05; thus, they enjoyed a good fit of items with the GDINA model in gender, BA and MA groups. To delve into this issue, the GDINA model had a very good fit of the model by its absolute model fit indices.

Skill Mastery or Non-Mastery Pattern Probabilities

Another output of the GDINA model is the skill probability pattern which depicted the probability of answering a certain item having different mastery or non-mastery attributes. Test takers were classified into 2^5=32 latent classes. In 2019, females had a higher probability to rate as non-masters in all multiple groups. In line with the logic of the first latent class, the last latent class also favored males in mastering all reading skills. On the contrary, in 2020, females gained a higher probability of mastery in gender and BA classifications, and males in MA group.

Comparing Skill Mastery Probabilities

Items and attributes parameters are estimated by the application of GDINA model for the gender, BA, and MA groups. Results display the participants' skill mastery probabilities in the reading test in 2019 and 2020. The differences obtained from gender classification in skill probabilities based on gender differences in 2019 showed which group has a higher/lower chance of mastery. Extracting explicit information (0.248), syntax (0.256), and vocabulary (0.274) for females, and extracting explicit information (0.238) and syntax (0.257) for males received lower probabilities in comparison to the other skills. Interestingly, except for extracting explicit information, all the other skills for males were uniformly higher than females.

As for BA GPA classification in 2019, vocabulary (0.345), syntax (0.241), and making inferences (0.334) for females, and extracting explicit information (0.274), and connecting and synthesizing (0.352) for males had lower probabilities in comparison to the other skills. It is worth noting that except for extracting explicit information and connecting and synthesizing, all the other skills for males were uniformly higher than females. For MA GPA classification in 2019, all skills (vocabulary, syntax, extracting explicit information, connecting and synthesizing, making inferences) for females enjoyed uniformly lower probabilities (0.436, 0.215, 0.287, 0.352, 0.218) in comparison to males' skills (0.453, 0.250, 0.323, 0.404, 0.325). Males uniformly scored higher than females as well. In gender classification in skill probabilities based on gender differences in 2020, females outperformed lower in all skills (vocabulary 0.261, syntax 0.298, extracting explicit information 0.319, connecting and synthesizing 0.196, making inferences 0.424) in comparison to males' skills (0.318, 0.412, 0.614, 0.375, 0.575), respectively. And, males uniformly scored higher than females. For BA GPA classification in 2020, females did not outperform in vocabulary 0.177, extracting explicit information 0.276, connecting and synthesizing 0.166 and making inferences 0.383. In addition, except for syntax 0.255, all the other skills for males were uniformly higher (0.372, 0.410, 0.344, 0.546) than females. And, in MA GPA classification in 2020, males scored lower levels of skill probabilities (0.253, 0.249, 0.303, 0.262, 0.373) in all skills (vocabulary, syntax, extracting explicit information, connecting and synthesizing, making inferences) than females (0.367, 0.518, 0.488, 0.301, 0.501). Interestingly, females uniformly outperformed males.

Differential Item Functioning

To test DIF, the Wald test was run to show whether a set of parameters is equal to some values (Hou et al., 2014). More specifically, de la Torre and Lee (2013) evaluated the fit of the model at item level. According to Armstrong (2014), by means of the Bonferroni method to detect DIF items under multiple groups in CDM, the adjusted *p*-value was improved. It is worth noting that adjusted *p*-values from Tables 3 to 8 are based on the Bonferroni correction.

Table 3

Gender DIF Detection through the Wald Statistic in 2019

Items	Wald Statistic	Adjusted P-Value
1	119.6093	0.000**
2	22.0743	0.047***
3	28.4097	0.004***
4	5.8262	1.000*
5	43.1718	0.000**
6	90.7460	0.000**
7	11.8754	0.183*
8	90.9332	0.000**
9	44.8592	0.000**
10	72.6964	0.000**

* Non-significant, ** Large, *** Negligible

Table 4

BA GPA DIF Detection through the Wald Statistic in 2019

Items	Wald Statistic	Adjusted P-Value
1	12.6842	1.000*
2	12.8094	1.000*
3	15.6778	0.472*
4	17.9663	0.012***
5	3.6142	1.000*
6	9.7341	1.000*
7	14.7322	0.052*
8	14.0257	1.000*
9	25.4334	0.0004***
10	24.7095	0.017***

Table 5

MA GPA DIF Detection through the Wald Statistic in 2019

Items	Wald Statistic	Adjusted P-Value
1	16.8737	0.314*
2	6.5917	1.000*
3	14.1641	0.775*
4	2.8071	1.000*
5	11.2783	0.236*
6	34.6506	0.0003***
7	11.8341	0.186*
8	16.4811	1.000*
9	9.0038	0.610*
10	21.3119	0.063*

In Table 3, *p*-value denotes the typical significance level for the Wald statistic except for items 4 and 7. The results of adjusted *p*-value through Bonferroni suspected large DIF for items 1, 3, 5, 6, 8, 9, and 10 in reading comprehension under the fitted multiple group GDINA model. In Table 4, the *p*-value indicated the typical significance level for the Wald statistic except for items 1, 2, 3, 5, 6, 7 and 8. The results of the adjusted *p*-value through Bonferroni showed that items 4, 9, and 10 in the reading comprehension subtest flagged negligible DIF under the fitted multiple group GDINA model. In Table 5, the *p*-value presented the typical significance level for the Wald statistic for item 6. The results of the adjusted *p*-value through Bonferroni depicted negligible DIF in just item 6 in the reading comprehension subtest under the fitted multiple group GDINA model. As expected, based on the results of multiple groups GDINA model, these items had uniform DIF and favored male test takers. However, the effect size measures for most of the test items were almost non-significant or negligible in all gender, BA and MA degrees multiple groups.

Table 6

Gender DIF Detection through the Wald Statistic in 2020

Items	Wald Statistic	Adjusted P-Value
1	47.3184	0.000**
2	12.9237	0.116*
3	38.5841	0.0001**
4	72.4914	0.000**
5	62.8181	0.000**
6	62.6416	0.000**
7	89.1041	0.000**
8	96.3284	0.000**
9	48.5755	0.000**
10	72.9978	0.000**

Table 7

BA GPA DIF Detection through the Wald Statistic in 2020

Items	Wald Statistic	Adjusted P-Value
1	63.1242	0.000**
2	29.8831	0.0001***
3	17.1194	0.2889*
4	108.9499	0.000**
5	34.4246	0.0003***
6	24.3726	0.0198***
7	98.4498	0.000**
8	126.4546	0.000**
9	61.4211	0.000**
10	49.0316	0.000**

Table 8

MA GPA DIF Detection through the Wald Statistic in 2020

Items	Wald Statistic	Adjusted P-Value
1	63.1242	0.000**
2	29.8831	0.0001***
3	17.1194	0.2889*
4	108.9499	0.000**
5	34.4246	0.0003***
6	24.3726	0.019***
7	98.4498	0.000**
8	126.4546	0.000**
9	61.4211	0.000**
10	49.0316	0.000**

In Table 6, the *p*-value was the typical significance level for the Wald statistic except for item 2. The results of the adjusted *p*-value through Bonferroni flagged those items 1, 3, 4, 5, 6, 7, 8, 9, and 10 in the reading comprehension subtest had almost large DIF under the fitted multiple group GDINA model. Interestingly, in Tables 7 and 8, the *p*-value indicated the typical significance level except for item 3. The results of the adjusted *p*-value through Bonferroni suspected showing large and negligible DIF in items 1, 2, 4, 5, 6, 7, 8, 9, and 10 in the reading comprehension subtest under the fitted multiple group GDINA model.

Differential Attribute Functioning

Milewski and Baron (2002) approach was employed to find DAF. Providing diagnostic profiles of individuals – both for females and males – is also possible through the Mantel-Haenszel method (Holland & Thayer, 1988) to analyze attributes in a reading comprehension test.

Table 9
Gender Mantel-Haenszel Statistic for Detecting DAF in 2019

Attributes/Skills	MH Chi-Square	P-Value	Effect Size
Vocabulary	1.5001	0.2207	-0.2875
Syntax	1.5320	0.2158	0.2714
Extracting Explicit Information	0.5562	0.4558	-0.1961
Connecting and Synthesizing	0.2341	0.6285	-0.1122
Making Inferences	1.0034	0.3165	0.2229

Table 10
BA GPA Mantel-Haenszel Statistic for detecting DAF in 2019

Attributes/Skills	MH Chi-Square	P-Value	Effect Size
Vocabulary	0.4633	0.4961	-0.1566
Syntax	0.0301	0.8623	0.0373
Extracting Explicit Information	4.2127	0.0401	0.5282
Connecting and Synthesizing	0.3423	0.5585	-0.1345
Making Inferences	0.5246	0.4689	-0.1595

Table 11
MA GPA Mantel-Haenszel Statistic for detecting DAF in 2019

Attributes/Skills	MH Chi-Square	P-Value	Effect Size
Vocabulary	0.8231	0.3643	0.2479
Syntax	0.0352	0.8512	0.0474
Extracting Explicit Information	0.0355	0.8507	-0.0573
Connecting and Synthesizing	0.0207	0.8857	0.0388
Making Inferences	1.0432	0.3071	-0.2662

Table 12
Gender Mantel-Haenszel Statistic for detecting DAF in 2020

Attributes/Skills	MH Chi-Square	P-Value	Effect Size
Vocabulary	0.4167	0.5186	0.1674
Syntax	1.1345	0.2868	0.4279
Extracting Explicit Information	1.5579	0.2120	-0.2800
Connecting and Synthesizing	0.0804	0.7767	0.1501
Making Inferences	0.0041	0.9488	-0.0279

Table 13

BA GPA Mantel-Haenszel Statistic for detecting DAF in 2020

Attributes/Skills	MH Chi-Square	P-Value	Effect Size
Vocabulary	12.4840	0.0004	-0.9012
Syntax	0.6276	0.4282	0.3024
Extracting Explicit Information	1.7004	0.1922	0.2813
Connecting and Synthesizing	5.8715	0.0154	1.2902
Making Inferences	0.2324	0.6298	0.2010

Table 14

MA GPA Mantel-Haenszel Statistic for detecting DAF in 2020

Attributes/Skills	MH Chi-Square	P-Value	Effect Size
Vocabulary	5.6752	0.0172	-0.7150
Syntax	1.6498	0.1990	-0.5945
Extracting Explicit Information	6.9399	0.0084	0.6705
Connecting and Synthesizing	1.5145	0.2185	0.7554
Making Inferences	0.5499	0.4583	-0.3545

In 2019 and 2020, the *p*-values related to the gender Mantel-Haenszel Chi Square statistic in each attribute separately showed that none of the five reading comprehension attributes had statistically significant DAF against females, and the magnitude of DAF was statistically non-significant for all skills. The *p*-values in BA Mantel-Haenszel Chi Square statistic of each attribute in 2019 flagged one out of the five reading comprehension attributes, that is extracting explicit information, which had statistically significant DAF against females, but the magnitude was statistically negligible. In 2020, one out of the five reading comprehension attributes – that is connecting and synthesizing – carried statistically significant DAF against females, with negligible statistical magnitude. The *p*-values related to the MA Mantel-Haenszel Chi Square statistic in each attribute displayed that none out of the five reading comprehension attributes had statistically significant DAF against females in 2019; however, vocabulary and extracting explicit information suspected negligible DAF, and the magnitude of DAF was statistically negligible for all skills in 2020.

Discussion

In language testing and assessment standardized high-stakes testing is highly considered since it has important consequences for test takers. As for the first research question, this study set out to tackle DAF and DIF in high-stakes nationwide PhD admission tests for a relatively large number of participants through cognitive diagnostic assessment under the GDINA model. Moreover, it is aimed to explore the relative contribution of this study in language assessment with regards to substantive DIF which can denote not only construct irrelevant variance under DIF detection but also “fairness check” (Roever, 2007, p. 184) in second language assessment. In this case, “students’ performance interpreted relatively to the other group” (Brown, 2005, p. 3). Accordingly, in a large-scale assessment, collecting validity evidence is the foundation of test development. However, this cornerstone is empirically narrowed down for tests under study, which is in stark contrast to high-stakes test development routines, hence the result of which would be statistically significant differences in gender differences, BA, and MA degrees’ performances.

In fact, the findings of the study suspected large DIF against female in both 2019 and 2020, and in BA and MA degrees' GPA in 2020. To justify its rationale, the Mantel-Haenszel as a conservative method in DAF detection was run to observe whether the existing DIF in these items were clearly a case of non-mastery in attributes. Interestingly, almost all items were not statistically significant as shown earlier. That is to say, the promotion of "authentic" (Brown, 2005, p. 231) higher order mental skills rather than reading comprehension attributes were of paramount importance since mental skills enhance higher order learning skills, and, in turn, facilitate higher order language proficiency (Liaw, 2007). In addition, to improve the language proficiency of test takers, it is crucial to reconsider the validity of high-stakes test. With regards to the second research question, due to the potent role of critical thinking in language assessment, application of robust statistical analysis was recommended to analyze discourse in reading comprehension passages to uncover underlying conveyed message in texts. Van Dijk (2004, p. 352) believed in "the way inequality is enacted, reproduced, and resisted by text." In line with that Wodak (2001) claimed that different social groups ill-represent or mis-present in various types of discourse. Examples of the aforesaid issue are the fact that the adopted reading comprehension passages for PhD nationwide admission tests were about archeology which is a major merely offered at public universities, not Islamic Azad University, to a small number of female candidates. These can indirectly convey some presuppositions while doing a test. This would neglect the fact that "procedural fairness can be said to require that all test takers be treated in essentially the same way, and that their performances be evaluated using the same rules" (Kane, 2010, p. 178). This argument apparently required some challenges in exploring the facts resided between the lines, which make it difficult for all test takers in replying to test items.

As stated earlier, different statistical analyses were adopted in the present study and the results brought into light the discursive sources of fairness through CDA (van Dijk, 2006) in reading comprehension passages. That is to say, the manipulation of passages was manifested in unfair test items through specific choice of grammar, lexicon, and text organization (Fairclough, 1995) which can entirely prevent the emergence of true underlying abilities among test takers. In addition, unfair goal setting of test designers affects the entire plan of the system including "curriculum design circles, materials development, and program evaluation" (Brown, 2005, p. 252).

Conclusion

This study was an attempt at diagnostic assessment in reading comprehension for PhD nationwide admission test to attain fair tests. To fill several knowledge gaps in the field of cognitive assessment, it is recommended to generalize this study to include reading comprehension with values and attributes other than ones addressed. Brown and Hudson (2002, p. 275) discuss "the need for giving feedback" to enhance the quality of curriculum developers' task. As a result, they can more concentrate on attributes drawbacks. In sum, the conducting of this study was not without limitations. The content of Q-matrix suffers from subjectivity, and insufficient standardized methods for Q-matrix construction as yet define, which may cause content irrelevant variable and construct under representation (Messick, 1995).

ORCID

https://orcid.org/0000-0002-4416-3317

https://orcid.org/0000-0002-7957-671X

Acknowledgments

Not applicable.

Funding

Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

Rights and Permissions

Open Access

References

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67–91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x

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

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

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics*, *25*, 133–150. https://doi.org/10.1017/S0267190505000073

Brown, H. D. (2004). Language assessment: Principles and classroom practices. Pearson Education.

Brown, J. D. (2005). *Testing in language program: A Comprehensive guide to English language Assessment.* McGraw-Hill College.

Brown, J. D., & Hudson, T. H. (2002). *Criterion-referenced language testing*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524803

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*(2), 119–157. https://psycnet.apa.org/doi/10.1191/026553298667688289

de La Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*(3), 163–183. https://doi.org/10.1177/0146621608320523

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., & 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. http://dx.doi.org/10.1111/jedm.12022

diBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *cognitively diagnostic assessment* (pp. 361–390). Lawrence Erlbaum.

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* (pp. 970–1030). North-Holland Publications. https://www.scirp.org/

Fairclough, N. (1995). *Critical discourse analysis: The critical study of language.* Longman.

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. https://doi.org/10.1177/0265532210364380

Goldsmith-Phillips, J. (1989). Word and context in reading development: A test of the interactive-compensatory hypothesis. *Journal of Educational Psychology, 81*(3), 299–305. https://psycnet.apa.org/doi/10.1037/0022-0663.81.3.299

Haagenars, J., & McCutcheon, A. (2002). *Applied latent class analysis*. Cambridge University Press.

Hartz, S. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with pract*ice. University Illinois at Urbana Champain.

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*, 191–210. https://link.springer.com/article/10.1007/s11336-008-9089-5

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Lawrence Erlbaum.

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.12036

Hughes, C., & Jones B. (1988). *Integrating thinking skills and processes into content instruction*. Presented to the 3rd Annual Conference, Association for Supervision and Curriculum Development, Boston.

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

Jang, E. E. (2008). A Review of cognitive diagnostic assessment for education: Theory and application. *International Journal of Testing*, *8*(3), 290–295.

Jang, E. E. (2009). Demystifying a Q-Matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly*, *6*(3), 210–238. https://psycnet.apa.org/doi/10.1080/15434300903071817

Jiao, H. (2009). Diagnostic classification models: Which one should I use? *Measurement: Interdisciplinary Research & Perspective*, *7*(1), 65–67. https://doi.org/10.1080/15366360902799869

Johnson, M., Lee, Y. S., Sachdeva, R. J., Zhang, J., Waldman, M., & Park, J. Y. (2013). *Examination of gender difference using the multiple groups DINA model*. Paper presented at the 2013 Annual Meeting of the National Council on Measurement in Education, San Francisco CA.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272. https://doi.org/10.1177/01466210122032064

Kane, M. (2010). Validity and fairness. *Language Testing*, 27(2), 177–182. https://doi.org/10.1177/0265532209349467

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

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). Wiley-Blackwell.

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. 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*, 391–409. https://doi.org/10.1177/0265532215590848

Liaw, M. L. (2007). Content-based reading and writing for critical thinking skills in an EFL context. *English Teaching & Learning*, *31*(2), 45–87.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Milewski, G. B., & Baron, P. A. (2002, April). *Extending DIF methods to inform aggregate report on cognitive skills*. Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans, Louisiana.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*(3), 315–333. https://www.psychologie-aktuell.com/fileadmin/download/

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the ‘‘two disciplines’’ problem: Linking theories of cognition and learning with assessment and instructional practices. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (pp. 307–353). American Educational Research Association.

Roever, C. (2007). DIF in the assessment of second language pragmatics. *Language Assessment Quarterly, 4*(2), 165–189. http://dx.doi.org/10.1080/15434300701375733

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

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement, 44*(4), 293–311. https://www.jstor.org/stable/20461865

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

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

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

Shahmirzadi, N. (2023). Validation of a language center placement test: Differential item functioning. *International Journal of Language Testing, 13(1),* 1–17. https://doi.org/*10.22034/IJLT.2022.336779.1151*

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

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16*(1), 92–111.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, *83*(10), 758–765. https://doi.org/10.1177/003172170208301010

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

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305. http://dx.doi.org/10.1037/1082-989X.11.3.287

Van Dijk, T. (2004). Critical discourse analysis. In D. Schiffrin, D. Tannen & H. E. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 352–371). Blackwell.

Van Dijk, T. A. (2006). Discourse and manipulation. *Discourse & Society, 17*(3), 359–383. https://doi.org/10.1177/0957926506060250

Von Davier, M. (2005). *A general diagnostic model applied to language testing data*. ETS Research Report. No. RR-05-16. Educational Testing Service.

Wodak, R. (2001). What CDA is about: A summary of its history, important concepts and its developments. In R. Wodak, & M. Meyer (Eds.), *Methods of critical discourse analysis* (pp. 1–13). Sage Publications.

Yi, Y. (2012). *Implementing a cognitive diagnostic assessment in an institutional test: A new networking model in language testing and experiment with a new psychometric model and task type*. (Unpublished doctoral dissertation) University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

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