Language Testing in Focus: An International Journal

www.EUROKD.com

ISSN: 2717-9087

2024 (10)

Simulation study: Seeking fairness in a high-stakes placement test

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

ABSTRACT

The primary objective of Cognitive Diagnostic Assessment (CDA) lies in problem solving processes to systematically categorize test takers in mastery/non-mastery groups. What is of concern is using some attributes in reading comprehension test item options which are suspected to bias, and is simulating unbiased test items in a reversed engineering attempt. To this end, 4200 PhD candidates who sat for the examination were randomly selected. A Q-matrix was constructed, and raw data were fed into the R-studio software. Two groups of female and male differences were considered for the Differential Item Functioning (DIF) detection based on the Deterministic Inputs, Noisy "and" Gate (DINA) model. Results of the simulation study revealed that in 100 times of item generations, items were non-significant in favor of gender bias. In conclusion, this study was an attempt to determine whether an item identified with DIF is a correct detection or failed to be identified in fine-grained details.

Keywords
Bias, Cognitive Diagnostic Assessment, Differential Item Functioning, Gender, Reading Comprehension

Received
15 May 2024
Received in revised form
10 August 2024
Accepted
25 August 2024

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

Introduction

Within recent decades, Cognitive Diagnostic Assessment (CDA) is typically used *for* learning (Jang, 2008), which aims at classifying individuals with item response patterns. Historically speaking, the term *diagnosis* was used in different senses, specifically to find what is wrong with someone or something. However, Jang (2005, p. 1) notes that it is cumbersome "to detect language competence and provide critical reasoning". With regard to the importance of CDA, Shohamy (2001) believes that test users could not take advantage of washback because a scoring report which is not use-oriented cannot prepare information in detail about test takers' progress.

https://doi.org/10.32038/ltf.2024.10.02

In practice, a cut-off point cannot stand as an accurate criterion to reveal skill mastery profiles of test takers. To address this issue, Kane (2001) calls for showing underlying cognitive diagnostic feedback in reading skills to lessen a probable tension between the standardization and precision of test scores and assess fine-grained detail information.

Taking the above into account, CDA has proven its credit compared to some preliminary methods in educational measurement that is Item Response Theory (IRT), because in IRT local independence and uni-dimensionality (Lord & Novick, 1968) are merely two factors which maximize reliability and validity. However, in CDA, multidimensional subscales delineate meaningful interpretations (Rupp & Templin, 2008). Moreover, the major drawbacks of IRT-based procedures are that they are not cost-effective in terms of computer time analyses and are sensitive to sample size and model-data fit. Also, the areas between Item Characteristics Curves (ICC) do not associate with test of significance (Shepard et al., 1981; Hambleton & Swaminathan, 1985). Moreover, IRT models are latent continuous, whereas CDA models present either dichotomous levels such as mastery or non-mastery, or polytomous levels in a rating variable range from poor performance to outstanding performance (de la Torre & Minchen, 2014).

In detail, CDA is employed to show a cognitive model that can measure attribute mastery profiles of test-takers in the fine-grained detail. Leighton and Gierl (2007) believe that CDA manifests "simplified description of human problem solving on standardized educational tasks, which helps to characterize the knowledge and skills students at different levels of learning have acquired, identify the potential of mastery, and facilitate the explanation and prediction of students' performance" (p. 6) in educational measurement through constructing a Q-matrix.

In diagnostic assessment, the ultimate goal is to simultaneously examine strengths and weaknesses, and not measure the overall ability. This would be possible through Q-matrix construction as the initial step. To facilitate the process, developing diagnostic inferences could be described in a vector α with some variables including $\alpha = (\alpha_1, \ldots, \alpha_k)$. Here α_k argues the examinees' true abilities as either indication of mastery or non-mastery of this skill. This can be measured by each item through information given in a Q-matrix (Tatsuoka, 1985). In Q-matrix development, there are multiple sources of evidence consisting of related literature and think-aloud verbal protocol analysis reports. Then, a new Q-matrix is developed by combining the refined entries, which can correspond with required attributes to respond to an item. Thenceforth, Q-matrix development is substantive evidence to check fitness of items. To do so, the next step, which is removing unfair test items through Differential Item Functioning (DIF), is possible.

Nowadays, DIF analysis broadens its scope to detect item bias in case of gender, ethnicity, language, age, and most importantly psychological processes in testing (Ellis & Raju, 2003). To identify and remove items with DIF, revising, removing, replacing DIF items or administering the same test to another group of test takers are suggested. To resolve discrepancies in DIF detection, CDA pays heed to designing and collecting data in a realistic mode (Tatsuoka, 1990), which tends to enhance the underlying ability of population under

study. Because DIF analysis can assist test developers to determine fair or unfair test items. This approach would be strengthened by estimating real data and simulating as a ground for removing bias, in order to compare response patterns with regard to fixed attributes used in real and simulated analysis. Simulation also predicts clustering of patterns and relationships between attributes in a meaningful way in order to reveal the true attribute states (Kerr & Chung, 2012). Regarding the literature, no study has conducted to reveal DIF in simulated data on reding comprehension. For example, Hemati and Baghaei, 2020; Ketabi, et al., 2021; Kunnan, et al. (2022), Ranjbaran and Alavi, 2017; Roohani Tonekaboni, et al., 2021; Shahmirzadi, 2023; Shahmirzadi and Marashi, 2023; Shahmirzadi, et al., 2020 a, b; Tabatabaee-Yazdi, et al., 2021 analyzed reading comprehension in both formative and summative assessment. Moreover, Javidanmehr and Anani Sarab (2019), Kim (2015), and Mirzaei, et al. (2020) revealed non-diagnostic high-stakes testing in retrofitting CDA studies. As a result, DIF detection in real and simulated data in a PhD test is considered. To meet this purpose, the following research questions are posed:

RQ$_1$: Is there any true negative rate in the *P* of simulated data through CDA in the PhD test?

RQ$_2$: Is there any true negative rate in the *P*-holm of simulated data through CDA in the PhD test?

Method

Participants in the Qualitative Study

To evaluate the reading attributes and develop a Q-matrix in the qualitative phase of the study, five PhD candidates including two males and three females majoring in Applied Linguistics participated. They were between 25 to 45 years of age. After a brief training session to train how to code item attribute relationships among the attributes, 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. Subsequently, eight professors of Applied Linguistics comprising two males and six females who had 20 to 30 years of English language teaching experience were invited. Having reviewed the reading questions, the professors verbalized their thoughts in an open-ended structured written interview to explain the importance of using attributes in each test item. As the last step, a refined coded scheme was used to develop a Q-matrix.

Participants in the Quantitative Study

The test takers from this pool of 12,096 candidates were reduced to 4,200 through simple random sampling with the application of SPSS (Statistical Package for Social Sciences). This was done to make item selection, estimation, and interpretation far from bias. The sample consisted of both genders mostly aged between 25 and 50 years. The discrepancy among the selected participants can be explained in terms of gender differences. Participants had to complete a national test. Therefore, data were collected from the nationwide PhD admission test administered by the National Organization for Educational Testing (NOET).

Instrumentation

Qualitative and quantitative instrumentation

To specify how correctly attributes in CDA were chosen, it is worth defining the attributes in the wide variety of sources such as content domain theories, item content analysis, and think-aloud verbal protocol analysis. To do so, the researcher used some reading comprehension passages, a section of general English tests of the PhD test, to conduct think-aloud verbal protocol analyses. For implementing think-aloud verbal protocol analyses and developing a Q-matrix, an instrument was developed in accordance with Gao and Rogers' (2011) guidelines and Jang's (2009) reading skills frameworks. For designing the structured written interview, some questions based on critical thinking dispositions (Hughes & Jones, 1988) were included in the open-ended section of the structured written interview.

To depict the borders of fair or bias developed tests in general and items, in particular, a content analysis of the reading comprehension passages was conducted. The adopted test for the present research tests consisted of two different reading passages with 10 items written in four option multiple-choice tests. The collected raw data from these reading comprehension tests were utilized to feed into the relevant software, R-studio, to estimate DIF, and simulate data.

Data Collection Procedure

The examination – under study here – was run by NOET. Candidates who sat for the exam needed to test both their content knowledge and general English language proficiency in two separate booklets. The test, in a four-option multiple choice high-stakes test format, was for candidates holding an MA in Applied Linguistics. They sought to pursue their studies for a PhD at university.

Formal Q-matrix Construction

A Q-matrix is a "cognitive design matrix that explicitly identifies the cognitive specification for each item" (de la Torre, 2009, p. 2) since the more fine-grained the attributes are, the richer the diagnosis of candidates' strengths and weaknesses will be.

Here, Five PhD candidates majoring in Applied Linguistics were recruited to participate in a think-aloud verbal protocol analysis to gather information about possible cognitive processes involved in responding to reading comprehension test items of the present study. The participants were two males and three females between the age range of 25 to 45. After a brief training session, each student read each passage in a retrospective think-aloud session, and recounted the processes they used among the five provided attributes in approximately 20 minutes for 10 reading comprehension questions. Following this session, a follow-up open-ended structured written interview based on critical thinking dispositions was conducted. In this stage, the participants could clarify their statements on each item or attribute critically. Next, the panel of professors described earlier was invited to examine the extent to which each reading attribute resides in per test item. Having reviewed the reading questions, the professors immediately verbalized their thoughts in an open-ended structured written interview independently. The importance of using attributes in each test item was explained. They identified the skill(s) for each item and made annotations about the evidence on which they

based their assessments. As the last step, the researcher read through the written reports line-by-line in order to understand the reading skills involved. It is assumed that the initial framework was confirmed by the diagnostic analyses, such as raising their awareness in critical perception of the passage. However, care needs to be taken since participants' verbal reports mostly may not agree with expert rating (Jang, 2005; Zappe, 2007). Following the above procedure, a refined coded scheme was used to develop a Q-matrix.

Results

To assess DIF, the first step was developing a Q-matrix. To indicate desirable consensus among participants and experts' decisions, the Kappa Coefficient of Agreement was estimated. The coded Q-matrix of participants and experts was then fed into the SPSS software separately. There was substantial agreement between the two diagnoses, k=.78. Furthermore, the relationship between coded attributes was checked. The Phi Correlation Coefficient of Agreement showed inter-rater reliability. The results revealed that there was a negative 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 general, there were 4/15 (26.66%) strong correlation coefficients, 5/15 (33.33%) moderate correlation coefficients, and 6/15 (40%) weak correlations. Finally, the average correlation of a set of items that are indicative of the average correlation of all items estimated through Cronbach's Alpha reliability statistics was computed; the results displayed that there was almost a good reliability, α=.52. Table 1 shows the finalized Q-matrix.

Table 1

Developed Reading Comprehension Q-matrix for per Test Item

Items	Vocabulary	Syntax	Extracting Explicit Information	Connecting and Synthesizing	Making Inferences
1	1	0	1	0	1
2	1	0	0	1	1
3	1	0	1	0	1
4	1	0	1	0	0
5	0	1	0	1	0
6	1	1	0	1	0
7	0	0	0	1	1
8	1	0	1	1	1
9	0	0	1	0	1
10	1	0	1	0	1

As for checking the results of a CDA meaningfully and measuring the model fit indices, the other two main criteria consist of the Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarzer, 1978). The log–likelihood=-16245.3 led to the best-fitting model. Comparing AIC and BIC values, the lowest AIC=32595 and BIC=32924 values along with the best-fitting model for DINA were obtained. And, the mean of RMESA item fit was 0.01. In the next phase of analyses by the application of the R-Studio

package, DINA model used to test model fit indices and homogeneity between items, and detect item functioning orderly. The rationale for selecting this model was that the DINA model can clearly estimate slipping and guessing parameters and test each test item with assuming all the required number of attributes. Here, there was good fitness of items with the DINA model. A non-significant value of (Mx^2=142.18), p=0.00 indicated an acceptable model fit. Following the Mx^2 index, all other values also confirmed the DINA model fitted for the reading comprehension data as they were almost close to zero. The Standardized Root Mean Square Residual Index (SRMSR) for the DINA model indicated the best fitting model with data. There was the best fitting model, (SRMSR= .04), p<.05. Moreover, the results obtained from the absolute model fit indices revealed that values less than 0.25 were supposed to be tenable indices of the goodness of fit for individual items and the whole test. In MAD, all values approached zero which showed goodness of fit for multiple groups Generalized Deterministic Inputs, Noisy "and" Gate (GDINA model).

To estimate real data based on the DINA model, it was crucial to measure slipping and guessing parameters of both genders separately (Tables 2 & 3).

Table 2

Female Item Parameter

Items	Guess	Slip	IDI	RMSEA
I001	0.078	0.573	0.349	0.035
I002	0.030	0.780	0.191	0.004
I003	0.032	0.751	0.217	0.004
I004	0.042	0.747	0.211	0.012
I005	0.052	0.426	0.522	0.002
I006	0.032	0.759	0.208	0.003
I007	0.041	0.827	0.132	0.004
I008	0.076	0.784	0.139	0.007
I009	0.012	0.858	0.131	0.004
I010	0.094	0.733	0.174	0.011

Table 3

Male Item Parameter

Items	Guess	Slip	IDI	RMSEA
I001	0.085	0.403	0.512	0.055
I002	0.072	0.779	0.149	0.007
I003	0.053	0.666	0.282	0.013
I004	0.080	0.743	0.177	0.022
I005	0.063	0.215	0.722	0.003
I006	0.078	0.669	0.254	0.009
I007	0.041	0.780	0.179	0.014
I008	0.111	0.699	0.190	0.016
I009	0.040	0.852	0.108	0.012
I010	0.136	0.690	0.174	0.022

Next, two groups of female and male differences were considered for DIF detection based on the DINA model to compare real with simulated results. To estimate DIF, some commands were fed to run chi-square with the application of the difR package of R-studio (Table 4). The results suspected non-significant DIF in terms of gender bias.

Table 4

Gender DIF Detection through the Wald Statistic

Items	X^2	*P*	*P*-holm
1	0.3672	0.8323*	1.0000*
2	0.3165	0.8536*	1.0000*
3	0.4764	0.7881*	1.0000*
4	0.2491	0.8829*	1.0000*
5	1.2048	0.5475*	1.0000*
6	4.5175	0.1045*	0.9403*
7	0.9357	0.6264*	1.0000*
8	4.9897	0.0825*	0.8251*
9	0.0529	0.9739*	1.0000*
10	1.5258	0.4663*	1.0000*

Note: * Non-significant, ** Large, *** Negligible

Table 5

The Result of 100 Times Iteration of 10 Items in Gender Simulated DIF Detection

Items	Non-Significant Items in *P*	In Percent	Non-Significant Items in *P*-holm	In Percent
I001	1	99%	0	100%
I002	1	99%	0	100%
I003	1	99%	0	100%
I004	3	97%	0	100%
I005	0	0.0%	0	100%
I006	1	99%	0	100%
I007	0	0.0%	0	100%
I008	0	0.0%	0	100%
I009	0	0.0%	0	100%
I010	0	0.0%	0	100%

In Table 5, some arguments for real data analysis of all 10 test items were estimated based on the DINA model. It is noteworthy to add that slipping and guessing parameters which were obtained as preliminary fit indices in the previous stage were used in arguments related to simulation. The number of attributes for the simulated cognitive model was fixed at five and the sample size for per test item analysis was based on 4200 examinees. Subsequently, a set of commands for all 10 reading comprehension test items were simulated for 100 times item generations. At the last stage, the *P* and *P*-holm appearing in outputs were described. That is to say that all 10 items flagged no significant DIF where the null hypotheses were rejected. Interestingly, among 100 times of simulation, 93% of *P* and 100% of *P* holm were not significant. There was true negative rate in outputs of both *P* and *P*-holm. This could accentuate the power of simulated reading comprehension test items which transfer no bias. In other words, 7% *P* and 0% *P*-holm suspected false positive rate. In sum, the real data and simulated investigation were to a great extent conforming with each other. Because identifying items without DIF and failing to identify items with DIF, known as Type I error for the former and Type II error for the latter, did not occur.

Discussion

Simulation study is one of the most powerful tools in modern statistical analysis. In effective simulation, multiple modelling of components is possible. It is typically used by varying the

number of items and number of examinees, or by fixing the number of examinees and item parameters with regard to real data. Simulation predicts clustering of patterns and relationship between attributes in a meaningful way to reveal the true attribute states (Kerr & Chung, 2012).

In simulation study, showing clustering of items from different dimensions is done by Q-matrix development since the structure of items in Q-matrix can affect the accuracy of estimation in cognitive diagnostic procedure. Generally, there are many ways proposed in item generations; for instance, simulating 40 items, 7 attributes, and 1500 examinees, or 1500 examinees with various choices of randomly item generations with different attributes (Hartz, 2002). In sum, simulation studies are conducted to observe how well the true attributes reveal skill mastery profiles of test takers through item generations in robust statistical analysis. As a result, simulation study enhances the validity of test items by reducing Type I error.

In CDA, Q-matrix is developed to explain the relationship between attributes and items. Also, attempts have been made to discuss fairness and justice (Kane, 2010; Kunnan, 2010; Xi, 2010). Xi proposes fairness with respect to validity issues. Kunnan (2004, p. 33) believes “A test should not be harmful or detrimental to society.” Fairness also concerns equal treatment of the groups and individual test takers on the basis of avoiding psychometric bias to enhance reasonable and defensible judgments. What needs not to be ignored is that language testing community needs to be aware of the potential social harm derived from biased test items in tests in general and high-stakes tests in particular.

Generally speaking, tests should be relatively fair; and justice of the use of the test could be open to dispute as a result of which social values may change to add wider perspectives. This logic can be spread when higher level of mental skills and abstractions developed in a particular context, and resulted in renewing frameworks to syllabus design. Thus, designers need to consider expected test takers’ performances along with features of context (Wilson, et al., 2012). These properties demand for reasoning to design materials appropriate for particular purposes. To do so, test designers could take advantage of social and cognitive regularities which defend test validation procedures. To develop customized diagnostic assessment, assessment specialists and teachers also need to be involved in activities including specifications of specific skills to be learned, kinds of activities to facilitate skill development, and expected outcomes of learning. It is also essential to monitor the disciplines in which policymakers defined to avoid having unfair test items.

Conclusion

In this study, an attempt was made to show the skill mastery profiles of stake holders in a real and simulated educational context. This could provide useful information about test takers’ strengths and weaknesses in reading abilities unless test items were suspected of DIF. To this end, bridging the gap between theory and practice to reveal examinees interrelated but separable latent attributes mastery in reading comprehension test items could be possible. Accordingly, it would not be far-fetched to assert some theoretical and practical dilemmas which left some areas of research open; for example, this study can be extended to both different attributes and engage more participants of other majors so that more authentic

information would be available for developing fair tests in CDA. This study underwent reversed engineering analysis in CDA. Future research is needed to be engineered in designing a practical, useful, and in effect test which would conform to the standards of fairness. Selecting the best fit model other than the ones applied in the present study is suggested. De la Torre and Lee (2013) discuss the importance of selecting the best model fit although there would be some controversies with what was claimed as in practice, different results may be reported with choosing different models even in a condition that model fit indices were checked in advance.

To conclude, understanding of the fairness in the process of test development would lead to positive and widespread change. Many language testing communities empower each other to spread the vision of justice, fairness, and equ(al)ity in accomplishing their tasks rather than perceiving these concepts as some basic truths. Major expert test publishers should judge test contents in order to exclude unfair tests and include contents that reflects diversity of populations in responding to test items. In fact, the process of fairness review is pervasive to the extent that Ravitch (2003) describes it as “quietly endorsed and broadly implemented by textbook publishers, testing agencies, and professional associations” (p. 3). This may assist syllabus designers and materials developers to keep unsuitable materials from being generated. As a result, further investigations which might deepen understanding of the fairness in the process of test development would be beneficial; as this would lead into positive and widespread change. May language testing communities empower each other to spread the vision of justice, fairness, and equ(al)ity in accomplishing their tasks rather than perceiving these concepts as just some basic truths.

ORCID

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

Acknowledgments

My sincere gratitude is extended to Professor Parviz Birjandi, and Professor Hamid Marashi. To me, their expert guidance and trust have always been an inspiration.

Funding

Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

References

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

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., & 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

De la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa, 20*(2), 89–97. https://doi.org/10.1016/j.pse.2014.11.001

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).

Gao, L., & Rogers, W. T. (2011). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing, 28*(1), 77–104. https://doi.org/10.1177/0265532210364380

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. Boston: Kluwer-Nijhoff.

Hartz, S. M. (2002). A bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Hemati, S., & Baghaei, P. (2020). A cognitive diagnostic modelling analysis of the English reading comprehension section of the Iranian national university entrance examination. *International Journal of Language Testing, 10*(1), 11–32.

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*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Jang, E. E. (2008). A Review of cognitive diagnostic assessment for education: Theory and application. *International Journal of Testing, 8*(3), 290–295. https://doi.org/10.1080/15305050802262332

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing, 26*(1), 31–73. https://psycnet.apa.org/doi/10.1080/15434300903071817

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

Kane, M. T. (2001). So much remains the same: conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 53–88). London: Lawrence Erlbaum Associates.

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

Kerr, D., & Chung, G. K. (2012). Identifying key features of student performance in educational video games and simulations through cluster analysis. *JEDM-Journal of Educational Data Mining, 4*(1), 144–182. https://doi.org/10.5281/zenodo.3554647

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

Kim, A.Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2), 227–258. https://doi:10.1177/0265532214558457

Kunnan, A. J. (2004) Test fairness. In M. Milanovic & C. Weir (Eds.), *European Year of Languages Conference Papers*, Barcelona (pp. 27–48). Cambridge University Press.

Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing, 27*(2), 183–189. https://doi.org/10.1177/0265532209349468

Kunnan, A. J., Qin, C. Y., & Zhao, C. G. (2022). Developing a scenario-based English language assessment in an Asian university. *Language Assessment Quarterly, 19*(4), 368-393. https://doi.org/10.1080/15434303.2022.2073886

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26*(2), 3–16. https://psycnet.apa.org/doi/10.1017/CBO9780511611186

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

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

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

Ravitch, D. (2003). *The language police: How pressure groups restrict what students learn.* Knopf.

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.htm

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

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

Shepard, L., Camilli, G., & Averill, M. (1981). A comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, *6*(4), 317–375. https://doi.org/10.3102/10769986006004317

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., & Marashi, H. (2023). Cognitive diagnostic assessment of reading comprehension for high-stakes tests: using GDINA model. *Language Testing in Focus: An International Journal*, 8, 1–16. https://doi.org/10.32038/ltf.2023.08.01

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.

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.

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

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.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Lawrence Erlbaum.

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

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

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170. https://doi.org/10.1177/0265532209349465

Zappe, S. (2007). *Response process validation of equivalent test forms: How qualitative data can support the construct validity of multiple test forms*. Unpublished doctoral dissertation, Pennsylvania State University, State College, PA.