Tabaran Institute of Higher Education
ISSN 2476-5880
*International Journal of Language Testing*
Special Issue, 2023: *Advanced Psychometric Methods in Language Testing*

# Examining Attribute Relationship Using Diagnostic Classification Models: A Mini Review

Hajir Mahmood Ibrahim Alallo[*1], Aisha Mohammed[2], Zayad Khalaf Hamid[3], Aalaa Yaseen Hassan[4], Qasim Khlaif Kadhim[5]

**Abstract**
Diagnostic classification models (DCMs) have recently become very popular both for research purposes and for real testing endeavors for student assessment. A plethora of DCM models give researchers and practitioners a wide range of options for student diagnosis and classification. One intriguing option that some DCM models offer is the possibility of examining the nature of the interactions among the attributes underlying a skill. Attributes in second/foreign language (L2) may interact with each other in a compensatory/non-compensatory manner. Subskill/attribute relationship has been studied using diagnostic classification models. The present study provides a mini review of the DCM studies on the attribute relationships in L2 reading, listening, and writing. The criteria based on which interaction between the attributes have been inferred are reviewed. The results showed that the majority of DCM studies have investigated reading comprehension and more studies are required on the productive skills of writing and speaking. Furthermore, suggestions for future studies are provided.

*Keywords*: attribute relationship; Diagnostic Classification Models; DINA; DINO; GDINA

## 1. Introduction

For the ease of instruction and measurement, skills in second/foreign language (L2) have been broken down into subskills/attributes. The attributes can interact in a compensatory or non-compensatory manner in order to result in successful performance. When the relationship is compensatory, non-existence of a required subskill can be compensated by the presence of another subskill. This is the case when multiple strategies are available to solve the same task. On the other hand, when attributes interact with each other in a non-compensatory manner, all the attributes required for a task should be mastered/learned before one can solve the task successfully. Attribute relationships have been studied both theoretically and empirically (e.g., Gough & Tunmer 1986; Hoover & Gough, 1990). More recently, DCMs have been employed to study attribute

[1] English Department, College of Arts, Ahl Al Bayt University, Kerbala, Iraq. Email: hajaralbayati90@gmail.com
[2] College of Education, Al-Farahidi University, Baghdad, Iraq.
[3] English Department, AlNoor University College, Bartella, Iraq.
[4] English Department, Al-Nisour University College, Baghdad, Iraq.
[5] English Language Department, Al-Mustaqbal University College, Babylon, Iraq.

relationships. DCMs are suitable for studying interrationships between attributes since they assume either compensatory or noncompensaoty relationships between attributes.

*1.1 DCMs*

Cognitive diagnostic models (CDM) assign test takers to multidimensional skill profiles which classify them as masters and nonmasters of each of the subskills required by items in a test (Ravand et al., 2013; Ravand & Baghaei, 2020). For each dimension or subskill, CDMs provide mastery/nonmastery information for each test taker which, in turn, provides explanation for performance of each respondent on each item. Ravand (2016) grouped DCMs into general and specific. Specific DCMs assume either compensatory or non-compensatory relationships among the attributes required by items of a test while general models allow both types of relationships within the same test. The Deterministic Inputs, Noisy, “And” gate (DINA; Junker & Sijtsma (2001) is a noncompensatory model while the Deterministic, Input, Noisy “Or” gate (DINO; Templin & Henson, 2006), the Additive CDM (ACDM; de la Torre, 2011), linear logistic test model (LLM; Marris, 1999), and reduced reparametrized unified model (RRUM, Hartz, 2002) are compensatory. Generalized DINA (GDNA; de la Torre, 2011) and Loglinear CDM (Henson et al., 2009; LCDM) are general models.

*1.2 Attribute relationships using DCMs*

Application of DCMs to language testing data abounds (e.g., Hemati & Baghaei Moghadam, 2020; Ketabi, et al., 2021; Rohoohani, et al., 2021). To study attribute relationships using specific DCMs, one needs to apply several of these models. These studies are referred to in this paper as multi-DCM studies. With regard to the significance of multi-CDM studies, only a few studies have been conducted to compare several DCMs in order to identify the interaction of sub-skills in a particular cognitive domain. In their pioneering multi-DCM study on the listening and reading sections of iBT TOEFL (Test of English as a Foreign Language), Lee and Sawaki (2009) examined the functioning of three DCMs – the general diagnostic model (GDM; von Davier, 2008), latent class analysis model (LCA; Yamamoto, 1990), and fusion model or non-compensatory reparametrized unified model (NC-RUM; Hartz, 2002). They used a Q-matrix which was developed by Sawaki et al. (2009) based on the test specifications and content domain analysis by content experts. Then, NC-RUM was used to validate the Q-matrix. The analysis of model comparison showed a comparable performance of the three models with respect to examinee skill mastery classification, the accuracy of skill mastery probability distributions, and the test-retest reliability of examinee classification. Although Lee and Sawaki (2009) used the GDM as a general model which allows both compensatory and non-compensatory relationships between attributes, they only considered compensatory relationships between attributes of listening and reading.

Yi (2012) empirically compared the noisy input deterministic-or-gate (NIDO; Templin & Henson, 2006) model, deterministic-input noisy-and-gate (DINA; Junker & Sijtsma, 2001) model, deterministic-input noisy-or-gate (DINO; Templin & Henson, 2006) model, compensatory reparametrized unified model (C-RUM; Hartz, 2002), and the loglinear CDM (LCDM; Henson et al., 2009) on three sets of data, reading and listening of TOEFL and ECPE (Examination for the

Certificate of Proficiency in English), in terms of model fit and skill mastery profile patterns. To identify the required attributes, several domain experts were asked to brainstorm required attributes for each item. She used the NC-RUM to empirically validate the Q-matrix. Model comparison analysis showed that the performance of the C-RUM was similar to the LCDM. Due to the limitations of the output of the Mplus software, she had to use some of the results from Lee and Sawaki (2009). Findings of the study showed that the LCDM was the most optimal model with regard to both criteria and the C-RUM had the closest affinity to it. Overall, she concluded that the C-RUM as a compensatory model can better represent the relationships among sub-skills in language test data.

In another study, Li et al. (2015) compared the performance of multiple constrained models – two compensatory models (DINO, ACDM), and two non-compensatory models (DINA, RRUM) – against the G-DINA model on the reading section of the Michigan English Language Assessment Battery (MELAB). They based their comparisons on model fit and skill mastery profiles using the CDM package (Robitzsch et al., 2014) in R statistical software. They used a Q-matrix which was constructed by Li and Suen (2013) based on the review of related literature, students' think-aloud protocols, and expert judgment. The findings of model comparison indicated that compared to the GDINA model, the ACDM had the closest affinity to it followed by the RRUM, DINA, and DINO. They concluded that sub-skills underlying reading comprehension interact in a compensatory manner and should be modeled with ACDM.

Yi (2017a) compared the application of five CDMs (e.g., LCDM, DINA, DINO, NIDO, and C-RUM) to the scored response data of four forms of TOEFL reading and Listening comprehension sections in order to identify what they can show about the processing of L2 reading and listening attributes. Unlike previous studies in which only one attribute was defined for most items, Yi focused on the relative importance of attributes within and across items in L2 proficiency tests when several attributes are coded per item. The Q-matrices used in the study were taken from a research study by Sawaki et al. (2009) which includes a detailed content analysis of individual test items carried out by content experts. The Q-matrices were empirically analyzed using the fusion model to examine the appropriateness of the item coding. The models were compared in terms of model fit and item-correlation root mean squared errors (RMSE). It was found that the C-RUM had a better performance compared to the other rival models, and thus, the modeling scheme of the C-RUM can better reflect the processing of L2 listening and reading skills. In fact, L2 reading and listening attributes are compensatory.

The other multi-CDM study was conducted by Yi (2017b) who compared five CDMs, including LCDM, DINA, DINO, NIDO, and C-RUM, to find an optimal model for English as a second language (ESL) grammar test data. She used a Q-matrix which was previously constructed by Henson and Templin (2009) for the ECPE grammar based on Liao's (2007) work on basic factor structure of the grammar section of the Examination for the Certificate of Competency in English (ECCE; a similar test to ECPE. ECCE is suitable for intermediate learners, but ECPE is for advanced learners of English). Using exploratory factor analysis (EFA) and structural equation modeling (SEM), Liao explored three factors on which most ECCE items load (e.g., lexical knowledge, morpho-syntactic knowledge, and cohesive forms). Yi (2017 b) compared the models in terms of relative fit indices and RMSE. The results revealed that the LCDM and C-RUM had the best fit to the data relative to the other competing models. The results of model fit at item-level further showed that the C-RUM, DINA, and DINO were selected by six items with two attributes. Overall, Yi suggested that ESL grammar attributes interact in a compensatory manner. The presence of high correlations between grammar attributes supported their compensatory nature.

Yamaguchi and Okada (2018) made a comparison between item response theory (IRT) models and a wide range of CDMs subsumed by the GDINA model, including the DINA, DINO, ACDM, LLM, and RRUM, to examine which model has a better fit to the Trends in International Mathematics and Science Study (TIMSS) 2007 assessment data across seven countries. They employed a Q-matrix which was previously developed by Lee et al. (2011) based on the consensus of several domain expert researchers in mathematics education. The models were compared in terms of their absolute and relative fit statistics at test level. The results showed that the CDMs have better fit compared to the IRT models, and additive models (main effects models) showed better performance compared to the other CDMs. In other words, the results imply that there might be an additive relationship between attributes of TIMSS 2007 exam.

Aryadoust (2018) conducted a study to investigate the underlying structure of the listening test of the Singapore-Cambridge General Certificate of Education (GCE) exam and compare the functioning of six CDMs comprising GDINA, DINA, DINO, RRUM, multiple-choice DINA (MC-DINA; de la Torre, 2009; Ozaki, 2015), and higher order DINA (HO-DINA; de la Torre & Douglas, 2004) at test level to select the best model. To develop a Q-matrix, Aryadoust used three sources of information: listening comprehension theories and models, think aloud protocol analysis of test takers' test-taking strategies collected through interviews, and eye-tracking. Absolute and relative fit statistics were used to compare the six models. The results showed that the RRUM was the best model compared to the rival models, followed by the GDINA, and HO-DINA. Using the RRUM, Aryadoust examined the test-specific facets and cognitive skills of the GCE listening comprehension test. He concluded that there is a non-compensatory interaction between listening attributes. This finding was confirmed by the existence of a few negative and low correlations between listening attributes.

Ravand and Robitzsch (2018) compared the performance of a non-compensatory model (DINA), a compensatory model (DINO), three additive models (ACDM, C-RUM, and NC-RUM), and a saturated model (GDINA) on an Iranian high-stakes reading comprehension test. They initially criticized the previous multi-CDM studies in the following ways: (1) the major problem with Lee and Sawaki's (2009) study was that they used different software programs (e.g., HYBIL, mdltm, and Arpeggio for the LCA, GDM, and NC-RUM, respectively) to estimate the models. This practice did not allow the researchers to compare the functioning of the models based on the same fit statistics; (2) in previous studies, a non-compensatory model (e.g., NC-RUM) was used to validate a Q-matrix. They used the same Q-matrix to compare multiple CDMs. In this way, non-compensatory models were more likely to outperform their compensatory counterparts; (3) they only compared models at test level, and did not provide information at item level; and (4) they used small sample sizes for comparing models which could be insufficient for drawing a distinction among multiple CDMs (Lee & Sawaki, 2009). They used expert judgment to determine the attributes and then validated the Q-matrix empirically using the Q-matrix validation procedure developed by de la Torre and Chiu (2016) in the "GDINA" package (Ma & de la Torre, 2018) upon which a discrimination index can be used with all the constrained models subsumed by the GDINA model. They based their comparisons on model fit at both test and item levels, classification consistency and accuracy, and proportion of skill mastery profiles. The results of test-level analysis showed that the GDINA was the best model, and the C-RUM, NC-RUM, and the ACDM had the closest affinity to the GDINA. The DINA model showed comparable performance to the GDINA in terms of some criteria. Also, the item-level model comparison indicated that some of the multi-attribute items (e.g., items with more than one attribute) selected the DINA, DINO, and ACDM as the best fitting models. They concluded that the relationships among the attributes of reading comprehension might be a combination of compensatory and non-compensatory.

Another study was carried out by Du and Ma (2021) to diagnose a reading comprehension test designed by PELDiaG (Personalized English Learning: Diagnosis & Guidance) research team from a key university in China with the multi-CDM. The process of Q-matrix development in the study was informed by different sources: reviewing reading theories, experts' judgment, and students' verbal reports. They empirically validated the Q-matrix using the procedure suggested by de la Torre and Chiu (2016) in the "CDM" package (Robitzsch et al., 2018). Du and Ma (2021) compared the functioning of five CDMs (e.g., GDINA, DINA, DINO, ACDM, and RRUM) at both item and test level. The results of item-level analysis showed that different multi-attribute items picked different CDMs (e.g., DINO, ACDM, DINA, RRUM, and GDINA). They combined the five single CDMs and created a multi-CDM. At test level, the multi-CDM was compared against the single models. The models were compared in terms of absolute and relative fit indices, and the results indicated the better fit of the multi-CDM in comparison with the models. They also examined item profiles across the models. They concluded that a multi-CDM can better justify the inter-skill relationship between reading comprehension attributes, that is, a combination of compensatory and non-compensatory relationships can better explain the interactions of L2 reading comprehension attributes.

In a recent study, Shafipoor et al. (2021) compared model fit statistics, skill mastery probabilities, and classification accuracy produced by five constrained CDMs (DINA, DINO, ACDM, LLM, RRUM) against a general model (GDINA) as applied to the grammar and vocabulary sections of an Iranian General English Achievement Test, a truly diagnostic test. To develop a Q-matrix for the test, they consulted the literature on grammar and vocabulary theories and used expert judgment to determine the attributes. They then validated the Q-matrix empirically using the procedure developed by de la Torre and Chiu (2016). They also checked item-fit statistics, the mesa plot and the Heatmap plot for validating the Q-matrix. The analysis of model comparison at test level reveled the better fit of the GDINA and LLM relative to the other models. The LLM produced similar skill mastery proportions compared to the GDINA models. The results of classification accuracy also showed significantly high values for all of the models. At the item level, the analysis showed that multi-attribute items selected different models as their best fitting model. The findings of the study showed that the interactions among the attributes of vocabulary and grammar are a mixture of non-compensatory and compensatory.

As previous studies witness, most of the multi-CDM studies focused on receptive skills (e.g., reading and listening), and too little attention has been paid to the usefulness of multi-CDM studies in productive skills (e.g., speaking and writing). Recently, He et al. (2021) conducted a study to develop a diagnostic checklist using the descriptors of China's Standards of English Language Ability (CSE). Based on the descriptive parameters of CSE scale, required attributes were identified, and the relationship between the descriptors and were specified in a Q-matrix by four content expert. They first compared the performance of six CDMs comprising the GDINA, DINA, DINO, ACDM, LLM, RRUM at test level in terms of absolute and relative fit indices to find the best model. The results showed the better fit of the LLM compared to the rival models, indicating that the relationships among L2 writing attributes are compensatory. He et al. (2021) used the LLM to provide diagnostic feedback on test takers' attribute mastery profiles.

Although some studies have argued that reading comprehension attributes have non-compensatory relationships (e.g., Gough & Tunmer 1986; Hoover & Gough, 1990), most studies, be they CDM studies (e.g., Effatpanah et al., 2019; Li, Hunter, & Lei, 2015; Ravand & Robitzsch, 2018; Shafipur et al, 2021) or non-CDM studies (e.g., Goldsmith-Philips, 1989; Stanovich & West, 1979, 1981; Uso-Juan, 2006) point to the compensatory nature of the attributes underlying these reading comprehension skills.

Similarly, regarding the process of listening comprehension, several researchers have claimed that listening attributes interact in a compensatory manner. L2 listeners strategically use compensatory mechanisms (e.g., world knowledge, common sense, cultural information, and visual, contextual, or paralinguistic information) to compensate for their lack of knowledge in the target language (Vandergrift, 2007). According to Harding et al. (2015), "comprehension does not follow a strictly linear progression from the lower to the higher processing levels; rather, different levels may be operating concurrently, with breakdowns at one level compensated by "positive information" at another" (p.12). Thus, a number of researchers have suggested the use of compensatory models for exploring the relationship between listening attributes (Effatpanah, 2019; Yi, 2017a). On the other hand, there are other researchers who have paid special attention to the non-compensatory nature of listening and argue that listening attributes are highly interdependent (Vandergrift & Goh, 2012), and non-compensatory models can better reflect the interaction of listening attributes (Aryadoust, 2018; Buck & Tatsuoka, 1998; Sawaki et al.). In contrast, some researchers have argued that there is a combination of compensatory and non-compensatory relationships between listening attributes, and general models can better reflect the processing of L2 listening attributes (Meng, 2013; Dong et al., 2021).

## 2. Conclusion and suggestions for further research

Review of the literature shows that the majority of the studies have been carried out on reading comprehension and smaller numbers on listening with few on writing. The majority of these studies have compared performances of specific DCMs to decide how the underlying attributes interact. However, more recently some other studies (e.g., Ravand & Robtiztsch, 2018) rather than assuming the same relationships among all the items of a given test, have investigated item relationships at item level. Regardless of the way they have used to study interattribute relationships, almost all studies have found that in L2 skills, attributes interact with each other in a compensatory way. According to Ravand (2016) studying attribute relationships at item level might be more tenable because characteristics of each item may require different types of relationships for successful performance.

None of the above studies has explore relationships between attributes regardless of the responses to the items of the test. It is suggested that future studies study structural relationships between attributes using hierarchical DCMs. It is also suggested that future studies focus more on the productive skills of writing and speaking. Studies of measurement invariance within CDM are also encouraged (Ravand et al., 2020). Provision of feedback is more important when students engage in language production. Studies of the kind would be helpful to provide more light on the attributes of productive skills and how they interact with each other.

## References

Aryadoust, V. (2018). A cognitive diagnostic assessment study of the listening test of the Singapore–Cambridge general certificate of education O-level: Application of DINA, DINO, G-DINA, HO-DINA, and RRUM. *International Journal of Listening*, *35*(1), 29-52. https://doi.org/10.1080/10904018.2018.1500915

Du, W., & Ma, X. (2021). Probing what's behind the test score: Application of multi-CDM to diagnose EFL learners' reading performance. *Reading and Writing*, *34*, 1441-1466. https://doi.org/10.1007/s11145-021-10124-x

Effatpanah, F. (2019). Application of cognitive diagnostic models to the listening section of the International English Language Testing System (IELTS). *International Journal of Language Testing*, *9*(1), 1-28. URL: https://www.ijlt.ir/article_114295.html

Effatpanah, F., Baghaei, P., & Boori, A. (2019). Diagnosing EFL learners' writing ability: A Diagnostic Classification Modeling Analysis. *Language Testing in Asia, 9*:12. https://doi.org/10.1186/s40468-019-0090-y

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

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education,* 7, 6–10. doi:10.1177/074193258600700104

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilites: Blending theory with practicality*. (Doctoral dissertation). University of Illinois at UrbanaChampaign

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*. doi:10.1177/0265532214564505

He, L., Jiang, Z., & Min, Sh. (2021). Diagnosing writing ability using China's Standards of English Language Ability: Application of cognitive diagnosis models. *Assessing Writing*, *50*, 1-14. https://doi.org/10.1016/j.asw.2021.100565

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, 191–210. doi:10.1007/s11336-008-9089-5

Hemati, S. J., & Baghaei Moghadam, 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. Retrieved from http://www.ijlt.ir/article_114278_022ab6b64d4bc8acba9c561846828270.pdf

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*, 127–160. doi:10.1007/BF00401799

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258-272.

Ketabi, S., Alavi, S.M., Ravand, H. (2021). Diagnostic test construction: Insights from cognitive diagnostic modeling. *International Journal of Language Testing 11*(1), 1-143

Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing, 11*, 144-177.

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

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

Liao, (2007). Investigating the construct validity of the grammar and vocabulary section and the listening section of the ECCE: Lexico-Grammatical ability as a predictor of L2 listening ability. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *5*, 37–78.

Ma, W. & de la Torre, J. (2018). GDINA: The generalized DINA model framework. R package version 2.1. Retrieved from https://CRAN.R-project.org/package/GDINA

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187–212. doi:10.1007/BF02294535

Meng, Y. (2013). *Developing a Model of Cognitive diagnostic Assessment for College EFL Listening.* [Unpublished doctoral dissertation]. Shanghai International Studies University, China.

Ravand, H., Barati, H., & Widhiarso, W. (2013). Exploring diagnostic capacity of a high stakes reading comprehension test: A pedagogical demonstration. *Iranian Journal of Language Testing, 3, 11-37.*

Ravand, H., & Robitzsch, A. (2018). Cognitive diagnostic model of best choice: A study of reading comprehension. *Educational Psychology*, *38*(10), 1255-1277. https://doi.org/10.1080/01443410.2018.1489524

Ravand, H., Baghaei, P., & Doebler, P. (2020). Examining parameter invariance in a general diagnostic classification model. *Frontiers in Psychology*, 10:2930. Doi: 10.3389/fpsyg.2019.02930

Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: Recent development, practical issues and prospects. *International Journal of Testing*, *20*, 24-56. Doi: 10.1080/15305058.2019.1588278

Robitzsch, A., Kiefer, T., George, A., C., & Uenlue, A. (2018). CDM: Cognitive diagnosis modeling. R package version 6.3-45. Retrived from https://CRAN.R project.org/packageCDM

Roohani, 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), 1-143.

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

Shafipoor, M., Ravand, H. & Maftoon, P. (2021). Test-level and item-level model fit comparison of General vs. specific diagnostic classification models: A case of True DCM. *Language Testing in Asia, 11*(33)*,* 1-20. https://doi.org/10.1186/s40468-021-00148-z

Stanovich, K. E., & West, R. F. (1979). Mechanisms of sentence context effects in reading: Automatic activation and conscious attention. *Memory and Cognition*, 7, 77–85. doi:10.3758/BF03197588

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

Uso-Juan, E. (2006). The compensatory nature of discipline-related knowledge and English-language proficiency in reading English for academic purposes. *The Modern Language Journal*, 90, 210–227. doi:10.1111/j.1540-4781.2006.00393.x

Vandergrift, L. (2007). Recent Development in Second Language Listening Comprehension Research. In Graeme Porte (Ed.), *Language Teaching: Surveys and Studies*. (pp. 291-210) Cambridge University Press, Canada.http://dx.doi.org/10.1017/s0261444807004338 von Davier, M. (2008). A general diagnostic model applied to language testing data. British *Journal of Mathematical and Statistical Psychology*, 61, 287–307. doi:10.1348/000711007X193957

Yamaguchi, K., & Okada, K. (2018). Comparison among cognitive diagnostic models for the TIMSS 2007 fourth grade mathematics assessment. *PLoS One*, *13*(2), 1-17. https://doi.org/10.1371/journal.pone.0188691

Yamamoto, K., & Gitomer, D. H. (1990). Assessing knowledge structures in mathematics: A cognitive science perspective. In S. Legg & J. Algina (Eds.), *Cognitive assessment of language and math outcomes* (pp. 241{273). Norwood, NJ: Ablex Publishing.

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.

Yi, Y. (2017a). Probing the relative importance of different attributes in L2 reading andlistening comprehension items: An application of cognitive diagnostic models. *Language Testing*, *34*(3), 337-355. https://doi.org/10.1177/0265532216646141

Yi, Y. (2017b). In search of optimal cognitive diagnostic model(s) for ESL grammar test Data. *Applied Measurement in Education*, *30*(2), 82-101. https://doi.org/10.1080/08957347.2017.1283314