Language Testing in Asia

RESEARCH Open Access

Test-level and Item-level Model Fit Comparison of General vs. Specific Diagnostic Classification Models: A Case of True DCM

Mahdieh Shafipoor[1], Hamdollah Ravand[2*] and Parviz Maftoon[3]

* Correspondence: ravand@vru.ac.ir
[2]English Department, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran
Full list of author information is available at the end of the article

Abstract

The current study compared the model fit indices, skill mastery probabilities, and classification accuracy of six Diagnostic Classification Models (DCMs): a general model (G-DINA) against five specific models (LLM, RRUM, ACDM, DINA, and DINO). To do so, the response data to the grammar and vocabulary sections of a General English Achievement Test, designed specifically for cognitive diagnostic purposes from scratch, was analyzed. The results of the test-level-model fit values obtained strong evidence in supporting the G-DINA and LLM models possessing the best model fit. In addition, the ACDM and RRUM were almost very identical to that of the G-DINA. The value indices of the DINO and DINA models were very close to each other but larger than those of the G-DINA and LLM. The model fit was also investigated at the item level, and the results revealed that model selection should be performed at the item level rather than the test level, and most of the specific models might perform well for the test. The findings of this study suggested that the relationships among the attributes of grammar and vocabulary are not 'either-or' compensatory or non-compensatory but a combination of both.

Keywords: Attribute, General vs. specific diagnostic classification models, Model fit, Q-matrix, True diagnostic classification models

Introduction Diagnostic Classification Models (DCMs) are considered as paramount modeling alternatives for dealing with response data in the presence of multiple postulated latent skills, which can cause multivariate classifications of respondents (Rupp & Templin, 2008). These models have received much attention in the field of second language assessment in the last decade (Kim, 2015), and this interest is linked with increasing demands of discovering the learners' problems and finding a solution for them (Lee, 2015).

The main goals of DCMs are identifying strengths and weaknesses of individual learners to provide detailed feedback about their current knowledge and skills in order to take appropriate actions to remedy their weaknesses in various aspects of second

language ability (Lee & Sawaki, 2009a), and also classifying learners into similar skill mastery groups (Hartz SM: A Bayesian framework for the unified model for assessing cognitive abilities: blending theory with practicality, unpublished).

Different models with their statistical packages have been developed and applied so far. The Rule-space Model (RSM; Tatsuoka, 1983), Compensatory Reparametrized Unified Model (C-RUM, DiBello et al., 1995; Hartz SM: A Bayesian framework for the unified model for assessing cognitive abilities: blending theory with practicality, unpublished), Deterministic Input, Noisy "And" Gate Model (DINA; Junker & Sijtsma, 2001), Non-Compensatory Reparametrized Unified Model (NC-RUM; Hartz SM: A Bayesian framework for the unified model for assessing cognitive abilities: blending theory with practicality, unpublished), Deterministic Input, Noisy "Or" Gate Model (DINO; Templin & Henson, 2006), Attribute Hierarchy Method (AHM; Leighton et al., 2004), General Diagnostic Model (GDM; von Davier, 2005), Generalized DINA Model (G-DINA; de la Torre, 2011), Additive CDM (ACDM, de la Torre, 2011), and Hierarchical Diagnostic Classification Model (HDCM; Templin & Bradshaw, 2013) can be enumerated as some.

These models are differentiated by reflecting compensatory or non-compensatory relationships between postulated attributes of an item (Kunina-Habenicht et al., 2012). In a compensatory model, the mastery of one attribute can compensate for the lack of other attributes measured by the same item (de la Torre, 2011). On the contrary, when a model is non-compensatory, the mastery of all the attributes is required to answer an item correctly (de la Torre, 2011). The aforementioned relationships among the attributes of an item can have ramifications for model selection in DCMs. Since choosing the right model will make a difference in the classification of test-takers, it should be performed cautiously (Lee & Sawaki, 2009b).

There are also two directions in applying DCMs: (1) to use DCMs to develop true diagnostic tests from the onset and (2) to extract fine-grained diagnostic information from tests already developed for non-diagnostic purposes, a practice referred to as retrofitting (Lee & Sawaki, 2009b). Due to the fact that currently, explicit cognitive theories which underlie the design and development of educational assessments are missing, it may be a long time before truly diagnostic assessments can be developed (Liu et al., 2018). Therefore, except for a few attempts (e.g., Ketabi, S: Cognitive diagnostic analysis of reading comprehension: a case of undergraduate students' mastery over attributes across different fields of study, unpublished; Paulsen & Valdivia, 2021; Ranjbaran & Alavi, 2017) in which DCMs have been used to develop true diagnostic tests, most applications of DCMs (e.g., Aryadoust, 2018; Effatpanah, 2019; Jang, 2009; Li et al., 2015; Ravand and Robitzsch, 2018; Rupp & van Rijn, 2018; Yi, 2017) have taken the second line of action.

Another point conspicuously noticeable in the applications of DCMs in the field of language assessment is that they have mostly been applied to the reading comprehension tests (e.g., Hemati & Baghaei, 2020; Jang, 2009; Li, H: Evaluating language group differences in the subskills of reading using a cognitive diagnostic modeling and differential skill functioning approach, unpublished; Li et al., 2015; Ravand, 2016; Ravand et al., 2012; Ravand & Robitzsch, 2015, 2018), to a much lesser degree to listening comprehension (e.g., Aryadoust, 2018; Harding et al., 2015), sporadically to writing (e.g., Kim, 2011; Xie, 2016), much less sporadically to components of language such as

grammar (e.g., Lee & Sawaki, 2009a; Yi, 2017), and to the best knowledge of the authors, never to vocabulary.

Review of the literature

In the implications of DCMs, selecting a model among a large number of models is a difficult decision (Jiao, 2009). In most DCM studies conducted so far, one single DCM, indiscriminately, has been imposed on all items of the tests. There is a dearth of studies searching for the best DCM for different contexts. In addition, in view of the approaches applied to DCMs, the data extracted from the tests in these studies are either from an existing non-diagnostic test used for diagnostic purposes or a test developed from the beginning based on DCM guidelines for diagnostic determination. As a result, the studies relying on the selected approach are either retrofitting or true DCMs.

In a multi-DCM study, Lee and Sawaki (2009b) investigated the listening and reading sections of iBT TOEFL deploying one general model (GDM) and two non-compensatory constrained models (RUM, LCA). Although the findings revealed that all three models were comparable with regard to accurate test-takers' mastery classification and skill mastery probability, as well as a moderate across-form consistency, there was no outcome with common fit statistics due to utilizing three different software to run the models as it was only possible to use each software to estimate only one model. In addition, as no general model was applied against specific ones, the generalizability of the results to contexts in which both compensatory and non-compensatory interactions are allowed went under the question.

In another study, Yi (Yi Y: 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) compared one general model (LCDM) and four reduced models (ACDM, DINA, DINO, NDINO) in regard to model fit and skill mastery profiles. She found that the ACDM functioned similarly to the LCDM, but due to the MPLUS software limitations at the time of her study, it was impossible to come up with the results of fit indices for RRUM, a non-compensatory version of ACDM.

The software limitations of the two aforementioned model comparison studies in relation to showing different fit indices led to vague results about the performance of the models. To come up with software problems, Li et al. (2015), in a study, applied the R-package CDM, version 3.2-6 (Robitzsch et al., 2013). The CDM package in R is the most comprehensive software to run CDMs (Ravand & Robitzsch, 2018). Moreover, it makes the estimation of the fit indices of the other models possible (Lei & Li, 2014). To run the study, a general model (G-DINA), two reduced compensatory models (ACDM, DINO), and also two constrained non-compensatory models (RRUM, DINA) were applied to the Michigan English Language Assessment Battery reading test. The results of this study revealed that the ACDM possessed the closest affinity to the G-DINA model in view of the model fit and skill classification profiles; in contrast, the RRUM, DINA, and DINO models showed dissimilar results regarding both models fit statistics and skill mastery properties with the ACDM and G-DINA models.

All the above-cited studies applied different DCMs to EFL reading context with a focus on model fit and skill mastery profiles, but Yi (2017), in a new line of action, conducted more extensive research on multi-DCM by providing both basic statistics and

also evidence for validity and reliability. He compared the LCDM as a general model with the DINA, DINO, NDINO, and RRUM and reported that the RRUM model showed a better fit in a grammar test compared to the other models.

Still, in another study, Aryadoust (2018) investigated the fit indices of multi-DCM (G-DINA, DINA, HO-DINA, and RRUM) on the listening test of the Singapore-Cambridge General Certificate of Education exam using the CDM package in R and concluded that the RRUM model expressed the highest absolute fit indices in comparison with the other models. This result was in agreement with Yi (2017), who found the outperformance of RRUM compared with other models but incongruent with Chen and Chen (2016) and Ravand (2016), who reached the conclusions that the G-DINA model has the best-fit indices for a reading test. Since Lei and Li (2016) assigned a substantial role to the sample size in the performance of the fit indices and Galeshi and Skaggs (2014) emphasized the key role of the number of attributes measured by each item and the direct effects of the number of items and sample size on parameter recovery as well of classification accuracy, this controversy in Yi's (2017) findings might be as a result of existing many items with only one attribute in the test and the results reached by Aryadoust (2018) can be justified due to adopting a very small sample size (N= 205).

A recent study on multi-DCM was carried out by Ravand and Robitzsch (2018). The merits of this study over the other ones were applying a large sample size (N= 21, 42) to a high-stakes reading comprehension test comparing a General DCM (the G-DIND) with a reduced compensatory model (DINO), a constrained non-compensatory Model (DINA), and three additive models (ACDM, C-RUM, and NC-RRUM) in relation to both test-level and item-level fit indices, skill mastery profiles, and classification accuracy. To conduct the study, the CDM package in R was used. The results depicted that the G-DINA showed the best fit properties, and the C-RUM, NC-RUM, and ACDM performed almost the same as that of the G-DINA. Following the discrepancies in the results of multi-DCM studies in terms of the relationships among the attributes in a reading test, it was suggested that the DCMs should be run at the item level rather than the test level. As a consequence, each item can choose the model that best fits.

The studies on true multi-DCM are very few. In a very recent study, Ketabi (Ketabi, S: Cognitive diagnostic analysis of reading comprehension: a case of undergraduate students' mastery over attributes across different fields of study, unpublished) developed a true DCM reading comprehension test based on Ravand and Baghaei's (2019) framework comparing two general models, i.e., the G-DINA and LCDM, against four constrained models, i.e., the ACDM, the RRUM, the DINA, and the DINO in terms of model fit, proportions of different attribute probabilities, classification accuracy, and consistency among two groups of undergraduate students of humanities and engineering. The results showed that the ACDM would be the best model in terms of the model fit. In contrast, DINO model appeared to have the worst model fit. Furthermore, the students of humanities and engineering showed different attribute mastery. Moreover, this study did not estimate model fit indices at the item level to check the dependency among the items in order to remove the misfitting items and improve the model fit accordingly.

As mentioned above, there are a few studies on the DCM of best choice for reading (e.g., Jang, 2009; Li et al., 2015; Ravand & Robitzsch, 2018), less than a few on listening

(e.g., Aryadoust, 2018), and grammar (Yi, 2017) but no multi-DCM study on vocab yet. Another point worth noting is that, up to date, except for a few attempts (e.g., Paulsen & Valdivia, 2021; Ranjbaran & Alavi, 2017) in which DCMs have been deployed to develop true diagnostic tests, all the other existing multi-DCM studies have taken the retrofitting line of action (i.e., to use the existing non-diagnostic tests for diagnostic purposes). Hence, multi-DCM studies which apply DCMs to test development in order to provide diagnostic feedback are what is mostly required (Ravand & Robitzsch, 2018).

To top that off, the preeminent significance of the present study is applying a General English Achievement Test developed from the start based on DCMs. Furthermore, it is the first multi-DCM study on vocabulary. In addition, the present study investigated the performance of model fit indices similarly at the test and item levels, skill mastery profiles, and classification accuracy. Finally, the G-DINA package in R, version 3.6.3 (Ma & de la Torre, 2020) was performed to compare a general DCM model (G-DINA) against five specific models (LLM, RRUM, ACDM, DINA, and DINO) to decide on the model of the best choice for a General English Achievement Test. Although the CDM package is the most comprehensive software in R and can provide information on both relative and absolute fit indices to compare multi-models (Ravand & Robitzsch, 2015), the G-DINA package was used in this study since this software can handle different models in a unified manner which is not the case with the CDM package (Rupp & van Rijn, 2018). To add more, the G-DINA package is user-friendlier as it presents the results in both numerical and graphical formats, establishes easier interaction with the users, and is more time-efficient (Rupp & van Rijn, 2018).

Diagnostic Classification Models (DCMs) applied to this study

General DCMs (GDCMs) can assume different types of relationships in a test: compensatory, non-compensatory, additive, or hierarchical, and each item can decide on its own model. It is also plausible to use a GDCM due to its flexibility in allowing different kinds of interactions among the attributes when these interactions are blurred (Li et al., 2015). Examples of these models entail General Diagnostic Model (GDM; von Davier, 2005), Log-Linear CDM (LCDM; Henson et al., 2009), Generalized Deterministic Inputs, Noisy "And" Gate (GDINA; de la Torre, 2011), and Hierarchical Diagnostic Classification Model (HDCM; Templin & Bradshaw, 2013). The G-DINA model, which was applied to this study, in its saturated form, is not differentiated from other general models by relying on alternative link functions (de la Torre, 2011). It parametrizes both the main effects and the interaction effects of the attributes (de la Torre & Minchen, 2014), and if appropriate limitations are imposed, several specific DCMs can be derived from this general model (de la Torre, 2011).

According to the G-DINA model an item with two attributes $\alpha_1$and α_2 will possess one intercept parameter δ_{j0} (the probability of answering an item correctly without mastering any of the required attributes), two main effects $\delta_{j1\alpha+}\delta_{j2\alpha_2}$, and one interaction effect $\delta_{j12\alpha_1\alpha_2}$. In this case, the probability that test-taker *i*, answers item *j* correctly is expressed as follows (de la Torre, 2011):

$$\rho\left(x_j = 1|\alpha_1\ \alpha_2\right) = \delta_{j0+}\delta_{j1\alpha_1+}\delta_{j2\alpha_2+}\delta_{j12\alpha_1\alpha_2}$$

Specific (also called constrained or reduced) DCMs, contrary to the GDCMs, allow only one type of predetermined relationship in a test: compensatory, non-compensatory, additive, or hierarchical. Despite the fact that general CDMs are better at model-data fit, specific DCMs have less complicated interpretations and offer classifications with more accuracy (Ma et al., 2016).

For the purpose of this study, among different specific DCMs, a compensatory model: Deterministic Input Noisy "And" Gate Model (DINA; Junker & Sijtsma, 2001), a non-compensatory model: Deterministic Input noisy "Or" Gate Model (DINO; Templin & Henson, 2006) and also three additive models: Additive CDM (ACDM; de la Torre, 2011), Linear Logistic Model (LLM; Maris, 1999), and Non-Compensatory Reparametrized Unified Model (NC-RUM; Hartz SM: A Bayesian framework for the unified model for assessing cognitive abilities: blending theory with practicality, unpublished) were compared.

Considering the specific models, the DINA model can be derived from the G-DINA if the main effects and the interaction effects are set to zero (de la Torre, 2009, 2011), and it parametrizes slipping (s_j) and guessing (g_j) probabilities as well. Slipping probability is answering an item incorrectly despite having mastered all the attributes of that item, whereas guessing probability is responding to an item correctly even though the test-taker lacks the required attributes (Haertel, 1989). Thus, in this model, in case of a two-attribute item, the probability that test-taker i gets item j correct is as follows:

$$\rho\left(x_j = 1|\alpha_1\ \alpha_2\right) = {g_j}^{1-\alpha_1\alpha_2}\ \left(1-\mathrm{s_j}\right)^{\alpha_1\alpha_2}$$

The DINO model, similar to its compensatory version: the DINA has the guessing (g_j) and lack of slipping ($1 - s_j$) probabilities for item j. As the assumptions of the DINO model denotes, the probability of answering an item correctly is not differentiated if the test-taker has mastered all or even one of the required attributes for an item, and this probability is depicted as follows:

$$\rho\left(x_j = 1|\alpha_1\ \alpha_2\right) = {g_j}^{(1-\alpha_1)(1-\alpha_1)}\ \left(1-\mathrm{s_j}\right)^{1-(1-\alpha_1)(1-\alpha_1)}$$

The ACDM, RRUM, and LLM as additive models are derived from the G-IDINA by equating all the interaction effects to zero. In additive DCMs, each attribute has an additive role in increasing the probability of reaching a correct response, i.e., even the lack of one attribute can be compensated by other attributes (de la Torre, 2011). Another point worth noting is that by turning the identity link function in G-DINA to a log link function and a logit link function, the RRUM, and the LLM models can be run, respectively. Furthermore, the LLM model has a constant additive effect on the logit of the probability of a correct response (de la Torre, 2011).

To run the study, the research questions were posed as follows:

1) How do the G-DINA, DINA, DINO, ACDM, LLM, and RRUM models fit the grammar and vocabulary items of a General English Achievement Test at the test level?
2) How do the G-DINA, DINA, DINO, ACDM, LLM, and RRUM models fit the grammar and vocabulary items of a General English Achievement Test at the item level?

Method

Data sources

In the present study, a General English Achievement Test based on a cognitive framework was analyzed. The test was administered to 1773 male and female bachelor's university students taking part in a three-credit General English Language Course at Islamic Azad University, Shahr-e-Qdos Branch in 2019. They were studying different majors, and their age range was mostly 18 to 35. The test included two sections: grammar and vocabulary. Each section consisted of 30 four-option multiple-choice items. The allocated time to accomplish the test was 75 min.

The Achievement Test and the Q-matrix were adopted from Shafipoor (Shafipoor, M: A comparative study of different cognitive diagnostic models for developing a General English Achievement Test, unpublished), a recent study on true DCMs. Generally speaking, the framework for developing a cognitive diagnostic test mainly entails two approaches: Embreton's Cognitive Diagnostic System (CDS) (Embretson and Gorin, 2001) and Mislevy's Evidence-centered Design (ECD) (Mislevy, 1996). Although both approaches centralize the role of cognition and evidence gathering in test development, there are some differences in their procedures. There are some advantages of CDS approach to ECD approach, such as predicting item parameters for newly developing items, the possibility of learning about construct validity at both the test and item levels and gaining cognitive information, which enhances score interpretation (Embretson and Gorin, 2001).

Due to the advantages mentioned above, in Shafipoor's (Shafipoor M: A comparative study of different cognitive diagnostic models for developing a General English Achievement Test, unpublished) study, Embreston's CDS, along with the content analysis, students' think-aloud protocol, as well as experts' judgment were deployed to identify the attributes for test development purposes and decision making on the Q-matrix. As a matter of fact, in DCMs, the Q-matrix is a tentative vector and specifies the relationship between test items and the target attributes required by each item. Since the Q-matrix construction, to a large extend, is fulfilled by the experts in the field and is a subjective process, any misspecifications should be checked; otherwise, important practical implications may arise (de la Torre & Chiu, 2016).

In Shafipoor's (Shafipoor M: A comparative study of different cognitive diagnostic models for developing a General English Achievement Test, unpublished) study (See Appendix A), the Q-matrix consisted of four common attributes in both grammar and vocabulary sections of the test under the title of lexical skill, morphosyntactic skill, cohesive skill, and contextual meaning. The word "skill" encompassed both form and meaning. Technically speaking, the abovementioned attributes are defined by Purpura and E. (2004) as follows: lexical form refers to the ability to comprehend and produce the words which encode the grammar rather than meaning (e.g., syntactic features, co-occurrence restrictions); similarly, lexical meaning refers to the use of words and their interpretation through their literal meanings (e.g., collocations, false cognates, formulaic expressions). In addition, morphosyntactic form specifies understanding syntactic forms in a language (e.g., word order, syntactic structures); by the same token, morphosyntactic meaning ascribes meanings to syntax, inflections, or derivations (e.g., subjunctive mood, time/duration). Moreover, cohesive form refers to features of language which

make the interpretation of the cohesion at both sentence and discourse levels possible (e.g., referential forms, logical connectors/conclusions, adjacency pairs); likewise, cohesive meaning refers to the meanings conveyed through cohesive devices which connect the cohesive forms with their referential meanings within a context (e.g., personal referents, demonstratives, comparatives). Finally, contextual meaning assigns meaning to a message under the influence of interpersonal, situational, or social factors.

Data analysis

The G-DINA package, version 3.6.3 (Ma & de la Torre, 2020) in R software (R core team, 2013), was used to analyze the data comparing one general model (the G-DINIA) against five constrained models (DINA, DINO, ACDM, RRUM, and LLM).

As a result of identifying four attributes (K=4) involved in grammar and vocabulary items, sixteen latent classes (2^K) were recognized, and the proportions of the test-takers, classified in each of the attributes by different DCMs were estimated. After that, the data was analyzed in relation to model fit statistics at both the test and item levels.

Overall, at the test and item levels, fit indices are estimated at two levels: absolute fit indices, which compare the fit of the model to the data and relative fit indices, which compare the fit of a model with other rival models. Contrary to some researchers who are for absolute fit indices in multi-DCM studies (e.g., Li et al., 2015; Ravand, 2016; Yi, 2017), some others are against them (e.g., Chen et al., 2013; Lei & Li, 2016). The data analyses at the present study were conducted at both the test and item levels. First, the G-DINA model was compared against the fit of constrained models. Subsequently, model selection was performed at the item level to let each item choose the most fitting model.

Test-level model comparison

In the first phase of the data analysis, which was carried out at the test-level, three relative fit indices (AIC, BIC, -2LL), six absolute fit indices (M_2, $RMSEA_2$, SRMSR, proportion correct, log-odds ratio, transformed correlation), test-takers' skill mastery probabilities, and classification accuracy were estimated.

Akaike Information Criterion (AIC; Akaike, 1974), Bayesian Information Criterion (BIC; Schwarz, 1978), and −2 log-likelihood (-2LL) with small values count for a better data-model fit index. When there is a lack of Q-matrix misspecification, and a saturated model is applied, AIC depicts a higher accuracy compared to BIC; however, if larger sample size with less complicated models is used, BIC shows a better performance (Lei and Li, 2016).

M_2, $RMSEA_2$ (the root mean square error of approximation fit index for M_2), SRMS R (the standardized root mean squared residual), proportion correct (p), log-odds ratio (l), and transformed correlation (r) are the outputs of the absolute model-data fit indices of the G-DINA package.

M_2 is sensitive to local item dependency, misspecifications of the model, the Q-matrix, and the distributions of latent dimensions (Henson et al., 2009). A significant p value is the indication of the violation of the item independency and the misfit of the model to the data (Hu et al., 2016). $RMSEA_2$ "is a measure of discrepancy between the observed covariance matrix and model-implied covariance matrix per degree of freedom" (Chen, 2007, p. 467). $RMSEA_2$ ranges from 0 to 1 and values less than .06

indicate good fit (Hooper et al., 2008). SRMSR is a degree of the mean of standardized residuals between the predicted and the observed covariance matrices (Chen, 2007). The acceptable SRMSR values range between 0 and .08 (Hu & Bentler, 1999).

Proportion correct is the residuals between the predicted and observed proportions of test-takers' correct replies to the items. Log-odds ratio is the residuals between the predicted and observed logs-odds ratios of the item pairs. And transformed correlation refers to the residuals between the predicted and observed Fisher-transformed correlation of the item pairs. The smaller the values of all the abovementioned indices, the better the model fit will be.

Finally, classification accuracy is viewed as significant indices for evaluating the validity of classification results in DCMs (Wang et al., 2015) and is defined as the extent to which the test-takers' "classification of latent classes based on the observed item response patterns agrees with their true latent classes" (Cui et al., 2012, p. 23). Later, Iaconangelo (Iaconangelo, C: Uses of classification error probabilities in the three-step approach to estimating cognitive diagnosis models, unpublished) and Wang et al. (2015) introduced a new index for classification accuracy applying both pattern and attribute-level classification accuracy indices based on the G-DINA estimates for dichotomous data.

To evaluate the models in this study, classification accuracy at the test and attribute levels were checked. The evaluation was based on the G-DINA estimates followed the rules of Iaconangelo (Iaconangelo, C: Uses of classification error probabilities in the three-step approach to estimating cognitive diagnosis models, unpublished) and Wang et al. (2015).

Item-level model comparison

At item-level data analysis, the model fit was checked at the level of the item. There are few studies on item-level model fit (e.g., de la Torre & Lee, 2013; Henson et al., 2009; Ma et al., 2016; Sorrel et al., 2017), suggested different approaches (e.g., applying visual inspection, using the Wald test) to compare the model fit indices at the item level.

In the present study, the fit of the models at the item level was conducted following two approaches: in the first approach, the G-DINA model was run, and each item had the chance to select its best-fitting model. In the second line of action, the constrained models, suggested by the Wald test, were fitted to each single item following the rule of Ma et al. (2016), and the combinations of DCMs were compared against the G-DINA model.

Results

Q-matrix validation

To validate the initial Q-matrix, first, the suggested Q-matrix provided by the software was inspected by the experts. Since the suggestions were not in agreement with the experts' opinions, they were not considered.

Consequently, item-fit statistics, the Heatmap plot, the mesa plot, and the Item plots for each item were checked. The results showed there were some misfitting items. In order to identify and remove these items, dependencies between items were inspected

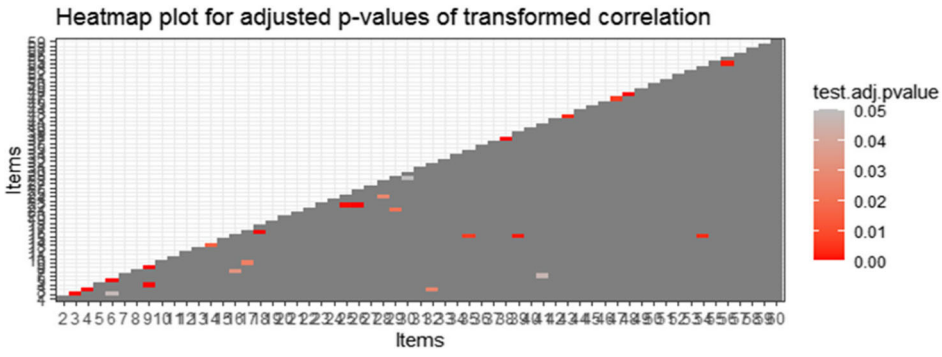

Fig. 1 Heatmap plot for the initial Q-matrix

through the values offered by transformed correlations and log odds ratio and, as a result, 22 items, i.e., 39, 28, 5, 38, 41, 25, 48, 32, 37, 8, 9, 3, 2, 4, 33, 35, 36, 43, 46, 18, 22, and 29, were removed which resulted in improved model fit. Tables 2 and 3 and Figs. 1 and 2 represent the findings.

As Table 1 shows, there was an improvement in both relative and absolute fit indices in the final Q-matrix compared to the initial one. For final Q-matrix, see appendix B.

Table 2 depicts that the adjusted *p* values for both transformed correlation and log odds ratio exceeded 0.05 and were insignificant.

Furthermore, the shading areas in Heatmap plot show the Bonferroni adjusted *p* values for all paired items. Red squares represent insufficient fit with *p* values below 0.05; in contrast, gray squares with *p* value above 0.05 indicate good fit indices. As illustrated in Figs. 1 and 2, by removing the misfitting items, the shading areas changed to gray.

Model comparison

Model comparison at the test level

Table 3 illustrates the relative and absolute fit values of the abovementioned models. In terms of the number of parameters, the G-DINA model with 175 parameters was the most complex one, while the DINA and DINO models possessing 91 parameters were the most parsimonious ones.

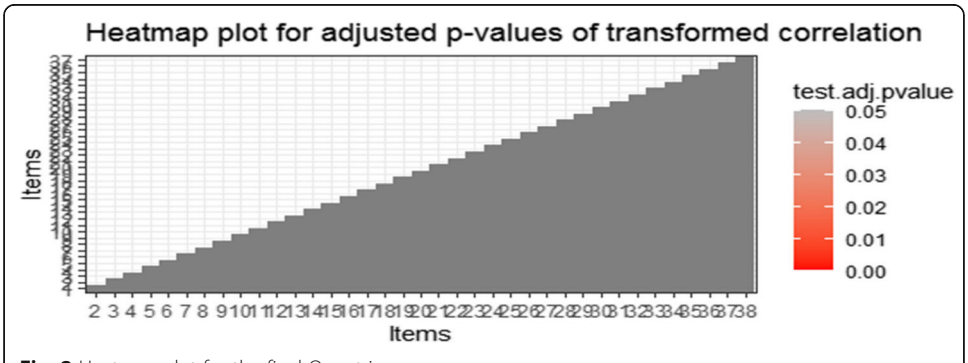

Fig. 2 Heatmap plot for the final Q-matrix

Table 1 Relative and absolute fit indices for initial and final Q-matrices

Model	#N par	AIC	BIC	M_2	$RMSEA_2$	$RMSEA_2$ CI 1	$RMSEA_2$ CI 2	SRMSR	-2log likelihood
Initial Q-matrix	267	120,336	121,799	3221	0.0245	0.0233	0.0257	.0399	119,802
Final Q-matrix	175	78,939	79,898	982	0.0204	0.0182	0.0225	0.0316	78,589

Note: #N par number of parameters, *CI* confidence interval

Considering the relative fit indices, the G-DINA model possessed the lowest AIC value, followed by LLM, ACDM, and RRUM. In addition, the LLM, RRUM, and ACDM had a very close affinity to the G-DINA. The DINA model was very similar to DINO in that it had the largest AIC values. In terms of BIC, the LLM had the lowest indices, followed by ACDM, RRUM, G-DINA, DINA, and DINO. The G-DINA did not show a low value for BIC. This can be probably justified by the sensitivity of BIC to highly parameterized models (Li et al., 2015).

In relation to the absolute fit indices, the ACDM, DINA, and DINO models yielded the largest M_2 and SRMSR values revealing the worst fit compared to other models; in contrast, the G-DINA indicated the best fit, followed by LLM and RRUM. The acceptable SRMSR values range between 0 and 0.08 (Hu & Bentler, 1999). As to $RMSEA_2$, although all the models obtained values <.05 and their confidence intervals in the upper bounds were <.05, the indices of ACDM and RRUM followed by LLM were identical to that of G-DINA and DINA. In comparison, DINO had the worst fit. $RMSEA_2$ ranges from 0 to 1 and values less than 0.06 indicate good fit (Hooper et al., 2008). In terms of -2LL, the G-DINA model obtained the smallest values. This was not far from the expectations since general models are highly parameterized, and they often show higher likelihood range compared to constrained models (Chen et al., 2013).

Briefly, as the results of the model fit comparison indicate, the G-DINA and LLM showed the smallest indices; then, they reached the status of the best-fitting models. The ACDM and RRUM were almost the closest models to the G-DINA. In addition, the value indices of the DINO and DINA models were very close to each other but larger than those of the G-DINA and LLM. Finally, the DINO appeared to have the worst model fit.

Table 4 illustrates the absolute item-level fit indices for each model. The results revealed that all models obtained good fit values to data. With respect to proportion correct values, the statistics in all the models were lower than the critical Z-score, i.e., 4.17. Contrarily, the indices of transformed correlations and log-odds ratios, except for the G-DINA model, were not satisfactory since their adjusted *p* values were lower than .05. This could be the case as a result of item dependencies in constrained models.

Table 2 Item-level fit indices for initial and final Q-matrices

		Mean[stats]	Max[stats]	Max[z.stats]	*p* value	Adj. *p* value
Initial Q-matrix	Proportion correct	00.01	0.0034	0.2907	0.7713	1
	Transformed correlation	0.313	0.2197	9.2443	0.0000	0
	Log odds ratio	0.1410	0.9237	8.6813	0.0000	0
Final Q-matrix	Proportion correct	0.0012	0.0041	0.3764	0.7066	1
	Transformed correlation	0.0258	0. 0928	3.9027	0.0001	0.669
	Log odds ratio	0.1142	0.4625	3.7646	0.0002	0.1173

Table 3 Model fit indices

Model	#N par	AIC	BIC	M_2	$RMSEA_2$	$RMSEA_2$ CI 1	$RMSEA_2$ CI 2	SRMSR	-2log likelihood
G-DINA	175	78939	79898	982.48	0.0204	0.0182	0.0225	0.0316	78589.31
LLM	131	79124	79842	1150.78	0.0224	0.0224	0.0204	0.0334	78862.88
ACDM	131	79229	79948	1132.69	0.0220	0.0200	0.0240	0.0355	78967.78
RRUM	131	79272	79990	1141.35	0.0222	0.0202	0.0241	0.0358	79010.32
DINA	91	80037	80536	1373.11	0.0250	0.0232	0.0269	0.0471	79855.69
DINO	91	80123	80622	1377.35	0.0251	0.0233	0.0270	0.0479	79941.99

Note. #N par number of parameters

Table 5 presents the observed pattern for all the models classified the test-takers into classes of mastery (1) or non-mastery (0). The most prevalent latent classes were class 1 [0000] with non-mastery of all the attributes and class 16 [1111] with mastery of all the attributes. With regard to the latent class indices, the LLM, ACDM, and RRUM were very identical to that of the G-DINA, while the DINA and DINO were the remotest.

In order to evaluate the agreement between the skill mastery probabilities of the G-DINA and the constrained models, the root mean square of the proportion difference (RMSPD) between the G-DINA model, and LLM, ACDM, RRUM, DINA, and DINO was calculated, and the values were 0.027, 0.028, 0.032, 0.053, and 0.078, respectively. As a result, the LLM followed by RRUM and ACDM had the closest affinity to that of the G-DINA, and the DINA and DINO were the most distant ones. In the next step, Cohen's Kappa as a method of checking the similarities between skill classification profiles of the G-DINA model and those of constrained models was computed. According

Table 4 Absolute item-level fit indices

Model		Mean[stats]	Max[stats]	Max[z.stats]	*p* value	Adj. *p* value
G-DINA	Proportion correct	0.0012	0.0041	0.3764	0.7066	1
	Transformed correlation	0.0258	0. 0928	3.9027	0.0001	0.669
	Log odds ratio	0.1142	0.4625	3.7646	0.0002	0.1173
LLM	Proportion correct	0.0013	0.0032	0.2732	0.7847	1
	Transformed correlation	0.0273	0.0983	4.1359	0.000	0.249
	Log odds ratio	0.1210	0.5072	3.9982	0.001	0.449
ACDM	Proportion correct	0.0018	0.0085	0.7535	0.4511	1
	Transformed correlation	0.0283	0.1117	4.6980	0.0000	0.0018
	Log odds ratio	0.1252	0.5093	4.5975	0.0000	0.0030
RRUM	Proportion correct	0.0012	0.0051	0.4499	0.6528	1
	Transformed correlation	0.0290	0.1169	4.9177	0.0000	0.0062
	Log odds ratio	0.1275	0.5283	4.7776	0.0000	0.0012
DINA	Proportion correct	0.0012	0.0030	0.2570	0.7972	1
	Transformed correlation	0.0390	0.1485	6.2483	0.0000	0
	Log odds ratio	0.1722	0.7906	5.6413	0.0000	0
DINO	Proportion correct	0.0012	0.0035	0.3092	0.7571	1
	Transformed correlation	0.00403	0.1603	6.7426	0.0000	0
	Log odds ratio	0.1787	0.8604	6.2667	0.0000	0

Table 5 Proportion of skill mastery profiles

Mastery pattern	G-DINA	LLM	ACDM	RRUM	DINA	DINO
0000	0.2700	0.2735	0.24636	0.26149	0.02992	0.4308
1000	0.0452	0.0500	0.04220	0.06752	0.03596	0.0539
0100	0.0155	0.0108	0.00287	0.00024	0.14829	0.0559
0010	0.1816	0.2041	0.20982	0.23740	0.13372	0.0055
0001	0.1951	0.1957	0.18055	0.16228	0.02929	0.0023
1100	0.0150	0.0316	0.00491	0.00099	0.03381	0.0345
1010	0.0979	0.0182	0.08362	0.09133	0.05535	0.0017
1001	0.0311	0.0179	0.00103	0.00935	0.06429	0.0424
0110	0.0157	0.0293	0.05331	0.06635	0.06307	0.0002
0101	0.0113	0.0065	0.00415	0.00252	0.00237	0.0176
0011	0.0016	0.0079	0.01406	0.01321	0.00181	0.0021
1110	0.0048	0.0261	0.01365	0.01936	0.04306	0.0542
1101	0.2199	0.0904	0.00858	0.00176	0.00047	0.0870
1011	0.0398	0.0578	0.07714	0.06381	0.04750	0.0509
0111	0.0124	0.0089	0.00727	0.00111	0.01634	0.0373
1111	0.1356	0.1917	0.23454	0.24054	0.39323	0.1229

to Cohen (1960), Cohen's Kappa indices are interpreted as slight agreement (< 20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), and almost perfect (0.81–0.100).

Table 6 indicates that the G-DINA and the LLM models reached the highest agreement, which ranged from

0.94 to 0.99 among different attributes. However, the DINO model had the least agreement with the G-DINA, and in relation to contextual meaning, the agreement was only 0.72 between these two models showing that 28% of the test takers were classified differently for this attribute. Also, the ACDM revealed high agreement with the G-DINA model followed by the RRUM, DINA, and DINO.

To compare the accuracy of the models, classification accuracy was calculated at test and attribute levels based on the G-DINA estimates applying approaches of Iaconangelo (Iaconangelo, C: Uses of classification error probabilities in the three-step approach to estimating cognitive diagnosis models, unpublished) and Wang et al. (2015).

Classification accuracy at the test level indicates the extent to which the test-takers are accurately classified into their true latent classes. Indices above .80 are acceptable rates for classification accuracy at the test level (Ravand and Robitzsch, 2018). As Table 7 indicates, the DINA, LLM, ACDM, and RRUM had identical accuracy indices to that of the G-DINA at the test level, but the only acceptable value was that of the G-DINA.

Table 6 Agreement of skill classification of the G-DINA and the constrained models

Attribute	G-DINA vs. LLM	G-DINA vs. ACDM	G-DINA vs. RRUM	G-DINA vs. DINA	G-DINA vs. DINO
Lexical skill	0.99	0.93	0.91	0.88	0.85
Morphosyntactic skill	0.98	0.89	0.87	0.75	0.75
Cohesive skill	0.96	0.93	0.91	0.80	0.87
Contextual meaning	0.94	0.90	0.89	0.88	0.72

Table 7 Test level accuracy

Model	G-DINA	LLM	ACDM	RRUM	DINA	DINO
	0.792	0.757	0.765	0.790	0.770	0.667

Classification accuracy at the attribute level shows the degree to which test-takers' are accurately classified into groups of maters or non-maters for each attribute. Johnson and Sinharay (2018) interpreted the classification accuracy as follows: values smaller than 20 represent lack of reliability, 0.25–0.50 poor reliability, 0.50–0.65 fair reliability, 0.65–0.80 good reliability, 0.80–0.90 very good reliability, and larger than 0.90 excellent reliability.

As Table 8 shows, in light of Johnson and Sinharay's (2018) rule of thumb, the values for the "lexical skill" showed excellent reliability in all the models. Furthermore, all the models classified test-takers accurately as masters and non-masters of every single attribute with the reliability values over 0.80, except for that of the DINA model with the least accurate attribute, i.e., "Morphosyntactic skill."

Model comparison at the item level

To check the fit of the models at the item level, two directions were applied. Following the first approach, the G-DINA model was run, and each item had the chance to pick its best-fitting model. Table 9 indicates the suggested models for items with multi attributes. Among 38 multi-attribute items, 17 items (i.e., 7, 13, 14, 15, 17, 20, 24, 27, 30, 31, 44, 49, 51, 54, 55, 57, 59) selected the LLM, 9 items (e.g., 6, 10, 16, 19, 26, 40, 47, 53, 60) picked the RRUM, 8 items (i.e., 1, 11, 12, 45, 50, 52, 56, 58) chose the ACDM, one item (i.e., 42) selected the DINA, one item (i.e., 21) took the DINO, and finally, two items (i.e., 23, 34) picked the G-DINA.

Next, in the second line of action, the fit of a new model (i.e., the combinations of DCMs applied to this study) was checked against that of the G-DINA model using the likelihood ratio test. Due to the fact that the new model involved constrained form of the G-DINA model, a smaller value of log-likelihood index and, consequently, better fit was expected. Table 10 shows the lower fit indices of the new model concerning both AIC and BIC values.

Discussion

The current study compared the model fit indices, skill mastery probabilities, and classification accuracy of a saturated model (G-DINA) against five constrained models (LLM, RRUM, ACDM, DINA, and DINO) with the grammar and vocabulary sections of a General Achievement English test designed specifically for cognitive diagnostic purposes from scratch.

Table 8 Attribute level accuracy

Attributes	G-DINA	LLM	ACDM	RRUM	DINA	DINO
Lexical skill	0.922	0.921	0.918	0.940	0.916	0.913
Morphosyntactic skill	0.915	0.936	0.919	0.925	0.735	0.879
Cohesive skill	0.851	0.861	0.934	0.951	0.832	0.858
Contextual meaning	0.922	0.927	0.926	0.927	0.901	0.873

Table 9 Model selection at the item level

Item	Model	*P* value	Item	Model	*P* value
1	RRUM	0.448	31	LLM	0.994
6	RRUM	0.9162	34	G-DINA	
7	LLM	0.214	40	RRUM	0.551
10	RRUM	0.737	42	DINA	0.695
11	ACDM	0.759	44	LLM	0.897
12	ACDM	0.433	45	ACDM	0.237
13	LLM	0.234	47	RRUM	0.367
14	LLM	0.909	49	LLM	0.982
15	LLM	0.109	50	ACDM	0.237
16	RRUM	0.215	51	LLM	0.378
17	LLM	0.979	52	ACDM	0.528
19	RRUM	0.458	53	RRUM	0.939
20	LLM	0.070	54	LLM	0.236
21	DINO	0.262	55	LLM	0.638
23	G-DINA		56	ACDM	0.501
24	LLM	0.823	57	LLM	0.768
26	RRUM	0.264	58	ACDM	0.585
27	LLM	0.757	59	LLM	0.678
30	LLM	0.598	60	RRUM	0.062

The results of test-level-model fit values obtained strong evidence in supporting the G-DINA and LLM models as showing the best model fit. The ACDM and RRUM were almost very identical to that of the G-DINA. The value indices of the DINO and DINA models were very close to each other but larger than those of the G-DINA and LLM. Finally, the DINO appeared to have the worst model fit. The outperformance of the G-DINA model in this study converges with Effatpanah (2019), Chen and Chen (2016), Ravand (2016), and Ravand and Robitzsch (2018) but in disagreement with Aryadoust (2018), Li et al. (2015), and Yi (Yi Y: 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; 2017). There are some points worth mentioning regarding the inconsistencies of the obtained results.

First, considering all the criteria in both relative and absolute fit indices at the test level, the G-DINA model did not show a low statistic for BIC. This result was congruent with Li et al. (2015) and can be probably due to the sensitivity of BIC to highly parameterized models.

Second, the scarce number of multi-DCM studies which compared the constrained models against the G-DIDA, even with gaining very similar results with that of the G-DINA, claimed the superiority of these models to the G-DINA due to using less number of parameters (e.g., Li et al., 2015; Yi Y: Implementing a cognitive diagnostic

Table 10 Likelihood ratio test for the combinations of DCMs with the G-DINA model

Model	LL	#N par	AIC	BIC	x^2	df	*P*
G-DINA	−39294	175	78589	78939	-	-	-
DCMs	−39321	131	78643	78905	53.71	44	<0.001

assessment in an institutional test: a new networking model in language testing and experiment with a new psychometric model and task type, unpublished, Yi, 2017).

Next, this study, in the same vein as some other studies (e.g., Galeshi & Skaggs, 2014; Ravand and Robitzsch, 2018; Yi, 2017), could not neglect the determining role of sample size, number of items and attributes, Q-matrix misspecifications, and also complexity of the models in modifying the outcomes of the model fit values.

As to the skill mastery proportions, the LLM, ACDM, and RRUM models were very similar to the G-DINA in terms of the latent class indices, while the DINA and DINO were the most distant ones. The most prevalent pattern for latent classes were class 1 [0000] and class 16 [1111]. In order to evaluate the agreement between the skill mastery probabilities of the G-DINA and the constrained models, both the root mean square of the proportion difference (RMSPD) and Kohen's Kappa between the G-DINA model and the LLM, ACDM, RRUM, DINA, and DINO models were computed. With respect to the root mean square of the proportion difference (RMSPD), the LLM followed by RRUM and ACDM had the closest affinity to the G-DINA, and the DINA and DINO were the most distant ones.

In relation to the Cohen's Kappa's agreement, the G-DINA and the LLM models reached the highest agreement. Also, the ACDM revealed high agreement with the G-DINA model followed by the RRUM, DINA, and DINO, while the DINO model had the least agreement with the G-DINA model.

Finally, classification accuracy indices of both general and reduced models at the test and attribute levels indicated significantly high and rather identical values to each other. High values of attribute-level accuracy might be due to applying a true DCM with items requiring multi attributes.

As a result, it can be concluded that even though the compensatory LLM model depicted very similar fit indices at the test level to that of the G-DINA, the fit indices for the non-compensatory RRUM model were not remote, either. Then, it is not far beyond the expectation to assume both compensatory/non-compensatory rather than "either-or" relationships for the attributes of grammar and vocabulary items in this study, and this result is incongruent with Yi's (2017) findings in which he came up with a compensatory relationship for the attributes of a grammar test.

Furthermore, as Rupp (2007) asserted, some DCMs could be sub-classifications of larger models which in its turn might question the distinction between the compensatory and non-compensatory relationships of these models.

The model fit was also investigated at the item level to check if applying a model to all the items was appropriate. To do so, two approaches were carried on. In the first approach, the G-DINA model was performed to let each item choose its best-fitting model. In this way, among 38 items, 17 selected the LLM and 9 picked the RRUM as the best-fitting models. The ACDM, G-DINA, DINA, and DINO models were selected by 8, 2, 1, and 1 items, respectively. In the second line of action, the suggested reduced models by the Wald test (a combination of DCMs) were fitted to each individual item. The results approved better fit indices for the combinations of DCM against that of the G-DINA model.

Thus, it can be concluded that model selection should be performed at the item level rather than the test level, and most of the constrained models might perform well for the test. Following the patterns of model selection by each item, out of 38 items, 34

items required 2 attributes, and only 3 items required three attributes. It is worth noting that none of the items with three attributes selected the G-DINA model. The justification that these items did not pick the general model might be the complexity or the unknown cognitive processes of the items (Henson et al., 2009). Moreover, item selection was performed with the priority given to the LLM and RRUM, followed by the ACDM.

Also, all the items with three attributes picked either the LLM or the RRUM model. This result is in contrast with the literature wherein large number of required attributes by an item led to the selection of a saturated model. This controversy can probably be due to the type of the DCM test applied in this study. As the test was an Achievement one developed based on a true DCM, the difficulty level and the complexity load of the items could be mitigated and the chance of misspecifications of the Q-matrix could probably be very slim. In addition, the large number of multi-attribute items might explain the logic behind the observed responses of the items.

Conclusion

Overall, concerning the results obtained from both test and the item level fit indices, the following implications are suggested. First, when the nature of the interactions among the attributes are blurred despite the drawbacks of the general models such as possessing more parameters, difficulty in estimation routines, and overfitting, performing a general model in which both compensatory and non-compensatory relationships of the attributes are possible, can be the best. Then, at the item level, the G-DINA model can be run to allow each item to select its best-fitting model since it is far beyond the expectation to look for a single reduced model to fit all the items in a test.

While the current study introduced practical implications in terms of model selection for a vocabulary and grammar Achievement Test developed based on the cognitive diagnostic principles from the beginning, it was limited in applying only one particular Q-matrix adopted from a study conducted by Shafipoor (Shafipoor, M: A comparative study of different cognitive diagnostic models for developing a General English Achievement Test, unpublished). Due to the subjective nature of the Q-matrix construction, even though with a rigorous validation process, multiple Q-matrices might be required and applied to compare the results.

Furthermore, developing a true DCM in this study could clarify the factors involved in the observed responses of the items, but there are still some other issues which are almost quintessential in model evaluation, such as the effect of sample size, person and item local independencies, grain size of the attributes, and the nature of the relationships between the attributes of an item which are required to be investigated.

Moreover, in this study, the G-DINA model was applied as a general model, and DINA, DINO, ACMA, LLM, and RRUM were performed as reduced models; other studies can be conducted to compare other general and constrained models.

Finally, this study attempted to take DCM into a classroom setting by developing an Achievement Test to look at the practical aspects of DCM in classroom settings. Other studies can be investigated by focusing on different skills, materials, and content to fill up the gaps between the theoretical and practical aspects of DCMs.

Abbreviations
DCM: Diagnostic Classification Model; C-RUM: Compensatory Reparametrized Unified Model; DINA: Deterministic Input, Noisy "And" Gate Model; NC-RUM: Non-Compensatory Reparametrized Unified Model; DINO: Deterministic Input, Noisy "Or" Gate Model; AHM: Attribute Hierarchy Method; GDM: General Diagnostic Model; G-DINA: Generalized DINA Model; ACDM: Additive CDM; HDCM: Hierarchical Diagnostic Classification Model; LCDM: Log-Linear CDM; RMSPD: Root mean square of the proportion difference; CI: Confidence interval

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40468-021-00148-z.

Additional file 1. Appendices

Acknowledgements
The authors thank all the university instructors and students took part in different phases of the study.

Authors' contributions
This manuscript was extracted from Mahdieh Shafipoor's dissertation. As a result, she handled the experiment, collected the raw data, and provided data analyses under the supervision of her dissertation supervisor, Dr. Hamdollah Ravand and her dissertation advisor, Professor Parviz Maftoon. The authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The datasets for the current study are available upon request.

Declarations

Competing interests
The authors declare that there are no competing interests.

Author details
[1]Department of English, Science and Research Branch, Islamic Azad University, Tehran, Iran. [2]English Department, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran. [3]Department of English, Science and Research Branch, Islamic Azad University, Tehran, Iran.

Received: 6 September 2021 Accepted: 4 November 2021
Published online: 25 November 2021

References

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Auto Control*, *19*(6), 716–723 https://doi.org/10.1109/TAC.1974.1100705.
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.
Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A multidisciplinary Journal, 14*(3), 464-504. https://doi.org/10.1080/10705510701301834
Chen, H. L., & Chen, J. S. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly*, *13*, 218–230 https://doi.org/10.1080/15434303.2016.1210610.
Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*, 123–140 https://doi.org/10.1111/j.1745-3984.2012.00185.x.
Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46 https://doi.org/10.1177/001316446002000104.
Cui, Y., Gierl, M. J., & Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*, 19–38 https://doi.org/10.1111/j.1745-3984.2011.00158.x.
de la Torre, J. (2009). DINA model and parameter estimation: a didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115–130 https://doi.org/10.3102/1076998607309474.
de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7.
de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253–273 https://doi.org/10.1007/s113360159467-8.
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*, 355–373 https://doi.org/10.1111/jedm.12022.
de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicolog'ia Educativa, 20*(2), 89-97. https://doi.org/10.1016/j.pse.2014.11.001
DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitive diagnostic assessment* (pp. 361-389). Erlbaum.
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.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*(4), 343–368 https://doi.org/10.1111/j.1745-3984.2001.tb01131.x.
Galeshi, R., & Skaggs, G. (2014). Traditional fit indices utility in new psychometric model: cognitive diagnostic model. *International Journal of Quantitative Research in Education, 2*(2), 113-132. https://doi.org/10.1504/ijqre.2014.064388
Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–321 https://doi.org/10.1111/j.1745-3984.1989.tb00336.x.
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, 32*(3), 317-336. https://doi.org/10.1177/0265532214564505.
Hemati, S. J., & Baghaei, P. (2020). A cognitive diagnostic modeling analysis of the English reading omprehension section of the Iranian national university entrance examination. *International Journal of Language Testing, 10*(1), 11–32.
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 https://doi.org/10.1007/s11336-008-9089-5.
Hooper, D., Coughlan, J., & Mullen, M. (2008, June). *Evaluating model fit: a synthesis of the structural equation modelling literature.* In 7th European Conference on research methodology for business and management studies (pp. 195-200).
Hu, J., Miller, M. D., Huggins-Manley, A. C., & Chen, Y. H. (2016). Evaluation of model fit in cognitive diagnosis models. *International Journal of Testing*, *16*(2), 119–141 https://doi.org/10.1080/15305058.2015.1133627.
Hu, L., & Bentler, P. M. (1999). Cut off criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), –55 https://doi.org/10.1080/10705519909540118.
Jang, E. E. (2009). Demystifying a Q-Matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly*, *6*(3), 210–238. https://doi.org/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. S., & Sinharay, S. (2018). Measures of agreement to assess attribute- level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement*, *55*(4), 635–664 https://doi.org/10.1111/jedm.12196.
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 https://doi.org/10.1177/01466210122032064.
Kim, A. Y. A. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: cognitive diagnostic assessment of L2 reading ability. *Language Testing*, *32*, 227–258 https://doi.org/10.1177/0265532214558457.
Kim, Y. H. (2011). Diagnosing EAP writing ability using the reduced parameterized unified model. *Language Testing, 28*, 509–541. https://doi.org/10.1177/0265532211400860.
Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement, 49*, 59-81. https://doi.org/10.1111/j.1745-3984.2011.00160.x
Lee, Y.-W. (2015). Diagnosing diagnostic language assessment. *Language Testing*, *32*(3), 299–316. https://doi.org/10.1177/0265532214565387.
Lee, Y.-W., & Sawaki, Y. (2009a). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly*, *6*, 239–263 https://doi.org/10.1080/15434300903079562.
Lee, Y.-W., & Sawaki, Y. (2009b). Cognitive diagnosis approaches to language assessment: an overview. *Language Assessment Quarterly*, *6*, 172–189 https://doi.org/10.1080/15434300902985108.
Lei, P. W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q-matrices. *Applied Psychological Measurement*, *40*(6), 1–13 https://doi.org/10.1177/0146621616647954.
Lei, P.-W., & Li, H. (2014). *Fit indices' performance in choosing cognitive diagnostic models and Q-matrices*. Philadelphia, PA: Paper presented at the annual meeting of the National Council on Measurement in Education (NCME).
Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: a variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*, 205–237 https://doi.org/10.1111/j.1745-3984.2004.tb01163.x.
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.
Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement, 78*(3), 357–383. https://doi.org/10.1177/0013164416685599
Ma, W., & de la Torre, J. (2020). G-DINA: an R package for cognitive diagnostic modeling. *Journal of Statistical Software, 93*(14), 1–26.
Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*(3), 200–217 https://doi.org/10.1177/0146621615621717.
Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212. https://doi.org/10.1007/bf02294535.
Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*(4), 379–416 https://doi.org/10.21236/ada291836.
Paulsen, J., & Valdivia, D. S. (2021). Examining cognitive diagnostic modeling in classroom assessment conditions. *The Journal of Experimental Education*, 1–18 https://doi.org/10.1080/00220973.2021.1891008.
Purpura, J., & E. (2004). *Assessing grammar*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733086.
R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/
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.org/10.1016/j.stueduc.2017.10.007.
Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, *34*(8), 782–799 https://doi.org/10.1177/0734282915623053.
Ravand, H., & Baghaei, P. (2019). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing*, *20*(1), 24–56 https://doi.org/10.1080/15305058.2019.1588278.

Ravand, H., Barati, H., & Widhiarso, W. (2012). Exploring diagnostic capacity of a high stakes reading comprehension test: a pedagogical demonstration. *Iranian Journal of Language Testing*, *3*, 11–37 https://doi.org/10.1177/0734282915623053.
Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research & Evaluation*, *20*, 1–12.
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.
Robitzsch, A., Kiefer, T., George, A., C., & Uenlue, A. (2013). *CDM: Cognitive diagnosis modeling. R package version 3.2-6.* https://CRAN.R-project.org/package=CDM
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*, 219–262 https://doi.org/10.1080/15366360802490866.
Rupp, A. A. (2007). *Unique characteristics of cognitive diagnosis models.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.
Rupp, A. A., & van Rijn, P. W. (2018). G-DINA and CDM packages in R. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 71-77. https://doi.org/10.1080/15366367.2018.1437243,
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136.
Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement*, *41*(8), 614–631 https://doi.org/10.1177/0146621617707510.
Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconception based on item response theory. *Journal of Education Measurement*, *20*, 345–354 https://doi.org/10.1111/j.1745-3984.1983.tb00212.x.
Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*(2), 251-275. https://doi.org/10.1007/s11336-013-9362-0
Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287–305 https://doi.org/10.1037/1082-989X.11.3.287.
von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Rep No. RR-05-16). ETS.
Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, *52*(4), 457–476 https://doi.org/10.1111/jedm.12096.
Xie, Q. (2016). Diagnosing university students' academic writing in English: is cognitive diagnostic modelling the way forward. *Educational Psychology, 37*(1), 26–47 https://doi.org/10.1080/01443410.2016.1202900.
Yi, Y. S. (2017). In search of optimal cognitive diagnostic model(s) for ESL grammar test data. *Applied Measurement in Education, 30*, 82-101. https://doi:10.1080/08957347.2017.1283314

Publisher's Note