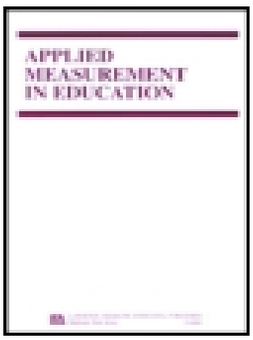


Applied Measurement in Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hame20

Detecting Differential Item Functioning Using Cognitive Diagnosis Models: Applications of the Wald Test and Likelihood Ratio Test in a University Entrance Examination

Roghayeh Mehrazmay, Behzad Ghonsooly & Jimmy de la Torre

To cite this article: Roghayeh Mehrazmay, Behzad Ghonsooly & Jimmy de la Torre (2021): Detecting Differential Item Functioning Using Cognitive Diagnosis Models: Applications of the Wald Test and Likelihood Ratio Test in a University Entrance Examination, Applied Measurement in Education, DOI: 10.1080/08957347.2021.1987906

To link to this article: https://doi.org/10.1080/08957347.2021.1987906

Published online: 13 Oct 2021.

APPLIED MEASUREMENT IN EDUCATION
https://doi.org/10.1080/08957347.2021.1987906

Routledge
Taylor & Francis Group

Detecting Differential Item Functioning Using Cognitive Diagnosis Models: Applications of the Wald Test and Likelihood Ratio Test in a University Entrance Examination

Roghayeh Mehrazmay[a], Behzad Ghonsooly[a], and Jimmy de la Torre[b]

[a]English Department, Ferdowsi University of Mashhad; [b]Faculty of Education, The University of Hong Kong

ABSTRACT

The present study aims to examine gender differential item functioning (DIF) in the reading comprehension section of a high stakes test using cognitive diagnosis models. Based on the multiple-group generalized deterministic, noisy "and" gate (MG G-DINA) model, the Wald test and likelihood ratio test are used to detect DIF. The flagged items are further inspected to find the attributes they measure, and the probabilities of correct response are checked across latent profiles to gain insights into the potential reasons for the occurrence of DIF. In addition, attribute and latent class prevalence are examined across males and females. The three items displaying large DIF involve three attributes, namely Vocabulary, Main Idea, and Details. The results indicate that females have lower probabilities of correct response across all latent profiles, and fewer females have mastered all the attributes. Moreover, the findings show that the same attribute mastery profiles are prevalent across genders. Finally, the results of the DIF analysis are used to select models that could replace the complex MG G-DINA without significant loss of information.

1. Introduction

Cognitive diagnosis models (CDMs) belong to a novel psychometric framework that assumes successful performance on a test item depends on the effective execution of the latent attributes measured by the item. Attributes, as defined by Birenbaum, Kelly, and Tatsuoka (1993, p. 443), include must-have "cognitive procedures, skills, or knowledge . . . to successfully complete the target task." By acknowledging the elicitation of more than one attribute by each test item, CDMs postulate multidimensionality within test items and delve into the examinees' mastery of each attribute required by items. CDMs serve as a tool to provide diagnostic feedback to examinees about the attributes they have or have not mastered, as in, their areas of strengths and weaknesses (Kim, 2015; Lee, de la Torre, & Park, 2012). As such, they can be contrasted with many item response theory (IRT) models that only provide a general ability estimate, θ, on a unidimensional scale.

CDMs have been either used as a framework to develop diagnostically informative tests or *retro*-fitted to tests not originally intended for diagnostic purposes (Gierl & Leighton, 2007). For the first purpose, the so-called cognitive diagnostic assessments are "expected to measure knowledge and processing skills with sufficient precision to make an examinee's incorrect response informative about what exactly the student does and does not know" (Leighton & Gierl, 2007, p. 148). To this end, all items are constructed in such a way that they provide maximal diagnostic information about the test takers' attribute mastery profiles. Regarding the second purpose, a particular CDM is retrofitted to an already administered test to provide "fine-grained diagnostic feedback beyond the aggregated test

CONTACT Roghayeh Mehrazmay Mehr1365@live.com Faculty of Letters and Humanities, English Department, Ferdowsi University of Mashhad, Azadi Square, Mashhad, Iran

scores" (Jang, 2009, p. 4). Analyzing extant tests to find the skills that result in successful performance on those items can provide researchers with insights into designing CDAs, and save them time and money (Lee & Sawaki, 2009).

A large number of CDMs have been developed, each of which is based on certain assumptions. Although CDMs can accommodate a variety of item responses and latent attributes (Chen & de la Torre, 2013; Rupp & Templin, 2008; von Davier, 2005; Yi, 2017), this paper focuses on dichotomous responses and attributes. In general, these models are divided into general (saturated) and reduced (constrained) CDMs. Reduced models can be derived from general CDMs by constraining certain parameters of the general models. Depending on the classification schemes, reduced CDMs can be classified as conjunctive or disjunctive, compensatory or noncompensatory, and additive or non-additive.

A compensatory model hypothesizes that nonmastery of an attribute can be compensated for, and result in successful performance on the item, by mastery of another required attribute (Li, Hunter, & Lei, 2015). As a rule of thumb, however, mastery of all the required skills maximizes the probability of a correct response to an item (Lee et al., 2012). In non-compensatory models, mastery of all the required skills is necessary for successful performance on the item, and non-mastery of any of the required attributes cannot be compensated for by mastery of other required attributes (Lissitz, Jiao, Li, Lee, & Kang, 2014). In other words, "the probability of a correct response is the same whether one or more than one of the required attributes is missed" (Li, 2008, p. 11).

Henson, Templin, and Willse (2009) distinguish between conjunctive and disjunctive noncompensatory models. They define conjunctive noncompensatory models in the same way as noncompensatory models were described above. Tatsuoka (1986) asserts that such models are typically suitable for math tests in which, as DiBello, Roussos, and Stout (2007, p. 1013) express, "the conjunction of successful skill execution on all the skills is required for success on the item." Describing such models as the opposite of conjunctive models, Henson et al. (2009) add that in disjunctive models, mastery of *any* subset of the required attributes results in successful performance on the item. As Roussos, DiBello, Henson, Jang, and Templin (2010) put it, disjunctive noncompensatory models are considered as 'extreme' compensatory models. Such models are more popular in psychology (Templin & Henson, 2006) or in education where multiple strategies can be used by the examinees to succeed on the item (Henson et al., 2009). Finally, additive models are those in which mastery of an attribute has a constant impact on the success probability expressed in various link functions (de la Torre, 2011).

Every test, cognitive diagnostic assessments included, aims at being fair to all test takers and ensuring invariance of measurement properties across examinees with distinct social, ethnic, or language backgrounds. Differential item functioning (DIF) analysis has long been used to ascertain measurement invariance with respect to examinees with diverse backgrounds. The issue of a DIF-free test is considered part of establishing the fairness and construct validity of the test (Sireci & Rios, 2013). DIF is present when two or more groups of examinees have dissimilar probabilities of answering an item correctly after being matched on their ability levels (Scheuneman & Bleistein, 1989). In other words, the ability level should account for all the differences in the examinees' performances; otherwise, the assumption of measurement invariance is violated (Millsap, 2011). In the context of CDM, the examinees are matched on their attribute mastery profiles, which are latent classes that indicate the attributes they have or have not mastered (Li, 2008). Under the CDM framework, uniform DIF is present when examinees from one particular group consistently have a higher or lower chance to endorse an item across all attribute mastery profiles. However, in non-uniform DIF, examinees who belong to a specific group have a higher chance of getting an item correct for some particular attribute mastery profiles and a lower probability of success for others (Hou, de la Torre, & Nandakumar, 2014).

Li and Wang (2015) argue that DIF results in contamination of item parameters and person attribute profiles, which, in turn, leads to the incomparability of latent classes across groups. Therefore, the validity and fairness of the whole test will be in jeopardy. Findings of a simulation study conducted by Wang, Guo, and Bian (2015) on the effect of DIF items on the accuracy of CDM

estimates illuminate the importance of performing DIF analyses. They reported that DIF had a more significant impact on the accuracy of the estimation of mastery profiles of the focal group than the reference group, and an increase in the magnitude and number of DIF items further reduced the estimation accuracy.

Despite its importance, only few DIF studies have previously been performed using CDMs. Utilizing the deterministic inputs, noisy "and" gate (DINA; Junker & Sijtsma, 2001) model, Zhang (2006) carried out a simulation study to compare the power of Mantel-Haenszel (MH; Mantel & Haenszel, 1959) and Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993) in identifying gender DIF. Results indicated that conditional upon the accuracy of the specified CDM and Q-matrix, matching on the attribute mastery profiles produced lower Type I error rates and higher power. Li (2008) applied a modified higher-order DINA (HO-DINA) model to a statewide mathematics test and concluded that MH performed better with larger sample sizes and more discriminating attributes. Hou et al. (2014) studied the capability of the Wald test in detecting DIF under the DINA model, and found that the Wald test exhibited smaller Type I error rates and greater power with an increase in the sample size, and outperformed both MH and SIBTEST in detecting uniform DIF.

Svetina et al.'s study (Svetina, Dai, & Wang, 2017) aimed at moving beyond simply finding DIF items by scrutinizing the patterns of skills and cognitive processes underlying items to explain the potential sources of DIF. MANOVA analyses indicated statistically significant differences among items regarding the patterns of elicited attributes, though with small effect sizes. In a more recent study, Ravand, Baghaei, and Doebler (2020) examined DIF across genders in the reading comprehension (RC) section of an English test in Iran using the Wald test and investigated the fit of multiple-group G-DINA (MG G-DINA). They reported no DIF in the test and also no significant difference between the two genders with regard to the correlations among the attributes and each gender's attribute mastery probabilities.

In summary, the very few applications of CDM DIF analyses to real data have mostly focused on identifying DIF items and have not delved into the sources of DIF. In addition, the benefits of using multiple-group models when there is more than one manifest group have been under-researched. To address the above-mentioned gaps, the present study aimed to:

(1) examine DIF across genders in the RC section of the Iranian University Entrance Examination (IUEE) and inspect the potential sources of DIF; and
(2) evaluate the improvement gained in the model due to using MG G-DINA, which allows separate parameter estimates for DIF items.

2. Theoretical Framework

2.1. Multiple-Group G-DINA

The G-DINA model, proposed by de la Torre (2011), is a generalized model from which some specific reduced models are derived. The general formulation of the model in identity function as presented by de la Torre (2011, p. 3) is as follows:

$$P\left(\alpha_{lj}^{*}\right) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{K'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \quad (1)$$

where
$\delta_{j0}$is the intercept for item j;
δ_{jk} is the main effect due to α_k;
$\delta_{j12}K_j$is the interaction effect due to α_1 and $\alpha_{K'}$; and
$\delta_j12\ldots_sK_j^*$ is the interaction effect due to $\alpha_1,\ldots,\alpha_{K_j^*}$.

For notational purposes, α_{lj}^{*} is used to represent the reduced attribute vector l containing only the subset of attributes required by item j, as opposed to all the attributes required by the whole test. In addition, the star symbol distinguishes the reduced attribute vector from the full one. The probability of a correct response equals the sum of the three elements present in the equation: a) δ_{j0} or the intercept for item j, which is equal to the guessing parameter for item j defined as the probability that an examinee who has not mastered any of the required attributes answers the item correctly through mere guessing (Zheng & Chang, 2016); b) the main effect due to mastering each individual attribute required by item j; and c) the interaction effects due to the mastery of two or more required attributes.

The G-DINA model yields different success probabilities for examinees mastering different numbers and combinations of attributes. Models such as the DINA, deterministic input, noisy "or" gate (DINO; Templin & Henson, 2006), additive CDM (A-CDM; de la Torre, 2011) model, linear logistic model (LLM; Maris, 1999), and reduced reparametrized unified model (R-RUM; Hartz, 2002) can be obtained from the above equation as special cases of the G-DINA model by constraining the G-DINA parameters (DINA, DINO, and A-CDM), and changing the link functions (R-RUM and LLM).

The G-DINA model implicitly assumes that all test-takers belong to the same population, an assumption that does not hold when test-takers come from various manifest groups as in studies of DIF, where two or more distinct groups of test-takers exist (George & Robitzsch, 2014). As an extension of G-DINA, the MG G-DINA model allows the item parameters and q-vectors to vary across groups for all or some items (Ma, Terzi, & de la Torre, 2020b). The probability of a correct response for an examinee belonging to the manifest group g in MG G-DINA is presented as:

$$P\left(\alpha_{glj}^{*}\right) = \delta_{gj0} + \sum_{k=1}^{K_{gj}^{*}} \delta_{gjk}\alpha_{glk} + \sum_{K'=k+1}^{K_{gj}^{*}} \sum_{k=1}^{K_{gj}^{*}-1} \delta_{gjkk'}\alpha_{glk}\alpha_{glk'} \ldots + \delta_{gj12\ldots K_{gj}^{*}} \prod_{k=1}^{K_{gj}^{*}} \alpha_{glk} \quad (2)$$

2.2. DIF in Cognitive Diagnostic Models

Under the CDM framework, DIF occurs when the reference and focal groups have different probabilities of answering the item correctly after being matched on their attribute mastery profiles. In other words, DIF is present if the probability of answering item j is different for the reference and focal groups in at least one of the attribute mastery profiles, shown as follows:

$$\Delta_{j\alpha l} = P(X_j = 1|\alpha_l)_{\mathrm{F}} - P(X_j = 1|\alpha_l)_{\mathrm{R}} \neq 0, \quad (3)$$

where $P(X_j = 1|\alpha_l)$ denotes success probability for examinees belonging to attribute mastery profile α_l on item j for either the reference (R) or the focal (F) group; and $\Delta_{j\alpha l}$,whose value determines which group the item is favoring, represents DIF in item j for the examinees. When $\Delta_{j\alpha l} < 0$ or > 0, item j favors the examinees in the reference or focal group with attribute pattern α_l, respectively.

In the context of CDMs, uniform DIF occurs when $\Delta_{j\alpha l}$ consistently has values greater or less than zero across all attribute mastery profiles. Non-uniform DIF, on the other hand, exists if $\Delta_{j\alpha l}$ yields positive values at certain attribute mastery profiles and negative values at certain others, a fact that signifies the reference group is advantaged only for some particular attribute mastery profiles.

As reviewed in the introduction section of the paper, previous studies have applied various methods to detect DIF. Below, the applications of the Wald test and the likelihood ratio test (LRT) to detect DIF are discussed based on the MG G-DINA model. To simplify notation, the following discussion is limited to two groups.

2.2.1. The Wald Test

The Wald test was first introduced into the CDM literature by de la Torre (2011), who used it at the item level to test whether a reduced model could replace the saturated G-DINA model with an acceptable fit. The Wald test was later used to detect DIF in the DINA model by Hou et al. (2014). This test uses multivariate hypothesis testing to detect DIF in the items.

The Wald test is implemented in two steps. In the first step, an unconstrained model is fitted to the data, and the item parameters are concurrently estimated by assuming the non-invariance of all item parameters across the focal and reference groups (Ma et al., 2020b). The item parameters are estimated in the form of a vector as shown below.

$$\begin{aligned}\hat{\boldsymbol{\beta}}_j^* &= \left(\hat{\boldsymbol{\beta}}_{Rj}, \hat{\boldsymbol{\beta}}_{Fj}\right)' \\ &= (\hat{P}\left(\alpha_{0j}^*\right)_R, \ \ldots\ , (\hat{P}\left(\alpha_{lj}^*\right)_R, \ \ldots\ , (\hat{P}\left(\alpha_{1j}^*\right)_R, \hat{P}\left(\alpha_{0j}^*\right)_F, \ \ldots\ , (\hat{P}\left(\alpha_{lj}^*\right)_F, \ \ldots\ , (\hat{P}\left(\alpha_{1j}^*\right)_F)\prime \end{aligned} \quad (4)$$

In the second step, the item parameters are constrained to be equal across the groups, and the model-data fit is checked to see whether the constrained model provides a worse fit to the data. To apply the constrained model in the second step, a restriction matrix of the dimensions $2^{K_j^*} \times 2^{K_j^*+1}$ is constructed where K_j^* is the reduced attribute vector that represents the number of attributes required by item j. To illustrate, if item j requires $K_j^* = 1$ attribute, a restriction matrix of 2×4 is created, as:

$$\boldsymbol{R}_j = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} \quad (5)$$

Given the restriction matrix, the Wald statistic to test DIF in item j is then computed as:

$$W_j = \left[\boldsymbol{R}_j \times \hat{\boldsymbol{\beta}}_j^*\right]' \left[\boldsymbol{R}_j \times Var\left(\hat{\boldsymbol{\beta}}_j^*\right) \times \boldsymbol{R}'_j\right]^{-1} \left[\boldsymbol{R}_j \times \hat{\boldsymbol{\beta}}_j^*\right] \quad (6)$$

where

R_j is the restriction matrix for item j;

$\widehat{\beta}_j^*$ is the estimated item parameters for item j stacked across groups; and

$\mathrm{Var}\left(\widehat{\beta}_j^*\right)$ is the variance-covariance matrix of the item parameters.

The tested null hypothesis for item j is $\mathrm{R_j} \times \widehat{\beta}_j^* = 0$, and the test statistic W_j is asymptotically chi-square distributed with degrees of freedom equal to $2^{K_j^*}$.

2.2.2. Likelihood Ratio Test (LRT)

To test for DIF, the LRT compares the likelihood of a reduced model to that of an augmented model. In the reduced model, the parameters of all the items are constrained to be equal across the focal and reference groups, whereas the augmented model permits the item parameters of the studied item to differ across groups while restricting the item parameters of the rest of the items to be equal. The LRT statistic (G^2) is, then, calculated as:

$$G^2 = -2\left[LL_{reduced} - LL_{augmented}\right] \quad (7)$$

where LL_{reduced} and $LL_{\text{augmented}}$ are the log-likelihoods of the reduced and augmented model, respectively. The LRT statistic is χ^2-distributed under the null hypothesis of no DIF, with degrees of freedom equal to the difference in the number of estimated parameters for the two models (Langer, 2008).

In contrast to the use of the LRT with IRT models, the application of LRT to CDMs does not require the identification of a set of unbiased items as the studied groups are on the same binary scale of mastery/non-mastery (de la Torre & Lee, 2010; Ma et al., 2020b). Ma et al. (2020b) propose a modified LRT under the CDMs framework in which the reduced model restricts the item parameters of the

studied item, thus, allowing the item parameters for the rest of the items to be estimated freely. In the augmented model, on the other hand, no restrictions are imposed on any items and the item parameters are freely estimated for all items across the reference and focal groups.

2.2.3. Effect Size Measure for the Magnitude of DIF

Zumbo (1999) suggests that a measure of DIF magnitude be provided alongside the test statistics to offset the effect of sample size and reduce the risk of false identification of DIF items. The present study utilizes two effect size measures, namely the unsigned area (UA) and the noncompensatory DIF (NCDIF), both of which are based on Raju's (1990) formulation of UA. The UA introduced by George and Robitzsch (2014, p. 414; Equations 8 & 9) calculates the absolute difference between the item response functions for the reference and focal groups for item *j*. They suggest classifying the magnitude of DIF as "moderate" and "large" if the effect size measure exceeds .059 and .088, respectively.

$$UA_j = \sum_{l=1}^{L} w(a_l)\left|P(X_j = 1\left|\alpha_l, g_1\right.) - P(X_j = 1\left|\alpha_l, g_2\right.)\right| \tag{8}$$

where

$$\mathrm{w}(a_l) = \frac{1}{2}\ [P(\alpha_l|G = g_1) + P(\alpha_l|G = g_2)]. \tag{9}$$

The second effect size measure is derived from the measure proposed by Oshima and Morris (2008, Equations 10 & 11). It is based on the average squared difference between the item response functions for the two groups for item *j*.

$$NCDIFj = E_F\left[d_j(a_l)^2\right] \tag{10}$$

where

$$d_j(a_l) = ES_{jF}(a_l) - ES_{jR}(a_l). \tag{11}$$

In the above equation, $ES_{jF}(a_l)$ and $ES_{jR}(a_l)$ are the probabilities of correctly answering item *j* for the focal and reference groups, respectively, conditional on the attribute vector a_l. The expectation of NCDIF for item *j* is taken over the attribute of the focal group. The cutoff values of .003 and .008 for NCDIF define moderate and large DIF (Wright & Oshima, 2015).

3. Method

3.1. Data and Participants

The present study utilized the data of the RC section of the IUEE, a high-stakes test taken by approximately one hundred thousand test-takers in Iran each year. The exam is administered annually to admit applicants seeking to study foreign languages into various undergraduate programs. IUEE comprises 70 questions in six parts, namely, Grammar, Vocabulary, Sentence Structure, Language Functions, Cloze Test, and RC section consists of three reading passages, each followed by five multiple-choice comprehension questions.

A sample of 12,169 candidates with an average age of 19.10 taking the test in 2017 was obtained from Iran's National Organization for Educational Testing. The data included responses of 4686 (38.5%) male and 7483 (61.5%) female candidates to the RC section. Furthermore, seven content experts participated in this study to design and validate the Q-matrix. To this end, two university professors, two Ph.D. graduates, and three Ph.D. candidates of teaching English as a foreign language (TEFL) with a mean of 11.2 years of experience in teaching reading comprehension courses were invited to identify the skills measured by each item.

3.2. Procedure

The first step toward conducting a CDMs analysis is to develop a Q-matrix (Tatsuoka, 1983) of the attributes the test items measure. The Q-matrix specifies the existing conceptual relationships between the items and the attributes they measure (Tatsuoka, 1986, 2009). There are several ways to construct the Q-matrix. Zhang (2006) states that Q-matrices are most often developed based on test blueprints or the judgments of the experts of the subject matter. Verbal reports such as think-aloud protocols can also provide hints on designing the Q-matrix (Leighton & Gierl, 2007). Previous research in the literature is also beneficial to better understand the cognitive processes involved in responding to items. However, the present researchers did not have access to the test blueprints; therefore, to develop the initial Q-matrix, relevant studies on RC and the experts' judgments were considered as ways of ascertaining the validity of the extracted attributes to the extent possible. To this end, previous research on RC was analyzed to come up with a list of attributes that RC tests generally aim to measure.

In a comprehensive review, Alderson (2000) summarized a large number of studies on reading tests and concluded that there was no unanimity among researchers as to, first, whether reading was a single unitary skill or divisible into a number of skills; and, second, what those skills were. He asserts that "the notion of skills and subskills in reading is enormously pervasive and influential, despite the lack of clear empirical justification" (Alderson, 2000, p. 10). Such a skills-based perspective serves CDM purposes well since such assessments are aimed at providing finer feedback about the test takers' weaknesses and strengths in the skills measured by the test. However, as cautioned by Alderson (2000), those lists lack an empirical basis. This claim is evidenced by the diverse list of attributes utilized in tests of RC (e.g., Cohen & Upton, 2006; Day & Park, 2005; Grabe, 2009; Jang, 2009; Jang, Dunlop, Park, & van der Boomet, 2015; Kim, 2015; Li & Suen, 2013; Ravand, 2016; Sawaki, Kim, & Gentile, 2009).

Given the disagreement on what skills RC tests measure, the researchers decided to examine the RC section of the IUEE to identify the skills measured by the test taking into account the attributes already identified in the previous studies. On one hand, the identified skills had to be broad enough so that there would be sufficient items measuring each skill; on the other hand, the list of the skills had to be detailed enough to provide fine-grained information about the test takers' RC abilities. The initial list of attributes included the seven skills of Vocabulary, Syntax, Reference, Details, Inference, Rhetorical Purpose, and Main Idea Comprehension. As the next step, the attributes were clearly defined based on similar attributes found in the literature, and the content experts were invited to identify the skills measured by each item to validate those skills. For each item, the experts were asked to put a "1" or "0" for each attribute if they believed the item was or was not measuring the attribute, or put a "?" if they could not decide one way or another. They were also asked to explain, name, or define any other attributes they believed the item measured, but were not provided on the list.

After the first round of coding was completed by the experts, the number of times each attribute was measured by the 15 RC items was calculated, and two of the attributes, namely Reference and Rhetorical Purpose, were found not to be measured by sufficient numbers of items. To address this problem, the experts suggested that the two attributes of Reference and Rhetorical Purpose be combined with the more general skills of Inference and Main Idea, respectively, on the basis of the features they shared with them. The coding was updated accordingly and the overall Fleiss Kappa agreement index (κ) among the experts was calculated. The average Fleiss Kappa index (κ) was .611, with a range from .545 to .679. Landis and Koch (1977) consider values between .41 to .60 and .61 to .80 as "moderate" and "substantial" agreement, respectively. The moderate agreement indices were obtained for Syntax and Vocabulary, where experts disagreed on the difficulty level of the vocabularies and grammatical structures of the passage. The initial Q-matrix and the values of κ for each attribute are reported in Appendix A.

Expert judgments per se are considered "fallible," the use of which introduces the danger of "a misspecified Q-matrix" (Chiu, 2013), which can, in turn, result in misclassification of the examinees as well as incorrect estimation of the model parameters (de la Torre & Chiu, 2016). Hence, the initially

developed Q-matrix was subjected to empirical validation in the GDINA package (Ma & de la Torre, 2020) in R (R Core Team, 2020). The GDINA package uses a discrimination index (ς_j^2) to identify for each item the most parsimonious q-vector that yields the highest variability in the probabilities of success among the latent classes. The q-vector with the fewest number of required attributes that accounts for a pre-specified proportion of variance (PVAF) relative to the maximum possible variance is suggested to replace the misspecified q-vector (de la Torre & Chiu, 2016).

To facilitate future applications of the GDINA package, generic codes of each step taken in the present study are provided in Appendix H. In the first run, a total of seven modifications in six items were suggested. The star symbol (*) next to the attributes in Appendix A denotes a suggested modification. The experts inspected the items for which there were suggested modifications, and considered the possibility of the inclusion of the specified attributes in the items' q-vectors or their exclusion therefrom. The experts agreed on applying modifications to items 1, 5, and 14 but insisted on keeping the original q-vectors for items 3, 7, and 10. The inspection of the mesa plots for those items helped further investigate the appropriateness of the specified q-vectors. As shown in Appendix B, the mesa plot for Item 3 depicts the original ("11001," in red bullet) and the suggested q-vector ("01001") as accounting for very similar PVAF values. Logically, the simpler q-vector is preferred. Nonetheless, the content experts emphasized the critical role of vocabulary for a correct response.

For Item 10, the original q-vector was simpler than the suggested one but accounted for less variance. Although the cutoff value is typically set to $\epsilon = .95$, Nájera, Sorrel, and Abad (2019) noted that the PVAF cutoff value may vary as a function of a number of factors, such as sample size, item quality, and test length. Thus, there is no hard and fast rule to interpret PVAF, and a value greater than .80, as for the original q-vector of Item 10, is still considered high enough. For Item 7, the interpretation of the mesa plot was not as straightforward. According to the mesa plot, the three vectors of "01000," "01010," and "01011" could be considered possible candidates, but the experts generally agreed that the item measures both Inference and Detail.

To check whether any further modifications were required, the revised Q-matrix served as the input for the second round of empirical Q-matrix validation. Four modifications were suggested for Items 3, 4, and 10 in the second run. Modifications to Items 3 and 10 were the same as those suggested in the first run; as for Item 4, it was suggested that Inference be dropped from the q-vector. When consulted, the experts unanimously opted to keep the original q-vectors for the three items. The mesa plots for Items 3 and 10, as previously discussed, and the one for Item 4 (see Appendix C), all favored keeping the original q-vectors. The more parsimonious q-vector (i.e., "01001") suggested for Item 4 offered a similar PVAF value. Nevertheless, the experts deemed Inference a necessary skill for this item. Since the experts did not agree with any of the suggested modifications, the Q-matrix obtained in the second round of revisions was selected as the final Q-matrix for subsequent analyses.

4. Results

4.1. Model Selection and Fit Analysis

The validated Q-matrix was used to select the model that best fitted the data. The GDINA package compares the fit of G-DINA as a saturated model to five reduced models, namely, the DINA, DINO, ACDM, RRUM, and LLM to find the best fitting model that can replace the saturated G-DINA model. The relative fit indices and the χ^2 tests comparing G-DINA to the reduced models all yielded significant *p*-values indicating that G-DINA could not be replaced by other reduced models. As shown in Appendix D, G-DINA had the largest log-likelihood and the smallest values for AIC and BIC. Furthermore, the LRT tests comparing the fit of G-DINA to each reduced model all yielded values greater than the critical values, resulting in significant *p*-values which meant the reduced models did not fit the data equally well.

Table 1. Absolute Fit Indices for the G-DINA Model.

Statistic	Value	
M2	5.791	$df = 5$ p-value = .327
RMSEA	.003	with 90% CI: [0, .0135]
SRMSR	.027	

Whereas relative fit indices help select the model that best fits the data among competing models, absolute fit indices determine whether or not the selected model adequately fits the data (Chen, de la Torre, & Zhang, 2013). Absolute fit indices are important to consider because they ascertain the validity of the inferences based on the CDM (Santos & de la Torre, 2019). Thus, the absolute fit indices for the G-DINA model, as provided in Tables 1 and 2, were checked in the next step. The GDINA package provides six fit indices, including the M2 statistic, RMSEA, SRMSR, proportion correct (PC), log odds ratio (LOR), and transformed correlation (TC). Nonsignificant values for M2, PC, LOR, and TC (de la Torre & Akbay, 2019; Hansen, Cai, Monroe, & Li, 2016) alongside values smaller than .05 for RMSEA and SRMSR (Maydeu-Olivares, 2013) indicate a good fit of the model. The M2 statistic for the G-DINA model was nonsignificant (p-value = .327). Both the RMSEA and SRMSR had values smaller than the recommended .05 value and the p-value for PC was nonsignificant, too. However, TC and LOR had significant p-values. Although there was no complete agreement among the different absolute fit indices, there was compelling evidence to support model-data fit.

A major concern about the RC test items is the issue of local or conditional dependence. As a central assumption of IRT models, local independence requires that a person's response to an item be independent from their response to other items in the test conditional on the person's ability (De Ayala, 2008; Stout, Nandakumar, & Habing, 1996). Statistically, this means that after conditioning on the latent trait, the correlation between residuals of item pairs should *not* be significant; otherwise, items are said to be locally dependent (Lee, 2004). A violation of local independence occurs "whenever the response process of one item provides the necessary cognitive schema to trigger a response to a subsequent item" (Ackerman, 1987, p. 1). Lee (2004) argues that RC items test related pieces of information provided in the passage which causes greater correlations among reading questions of a particular passage compared to correlations of items of a different passage. He adds that not all high inter-item correlations within the same passage are due to common latent trait, and "The extra item dependence within the passage, which is unaccounted for by this common latent trait, is very likely to lead to correlated residuals or LID [local item dependence]" (p. 78).

Within the context of CDMs, local independence has long been an inherent assumption (George & Robitzsch, 2014). Hansen et al. (2016) suggested applying Chen and Thissen's χ^2_{LD} index in the CDM context to complement overall fit analysis statistics such as M_2. The χ^2_{LD} statistic is used to detect local dependence among item pairs by evaluating the fit of the model to each pair of items. This statistic is provided in the output of the CDM R package (Robitzsch, Kiefer, George, & Ünlü, 2020) denoted as max(X2) accompanied by a test of significance with one degree of freedom. A non-significant value indicates the absence of local dependence. To test for the presence of local dependence in the present study, the max(X2) index obtained from the CDM package was checked. The max(X2) statistic was 232.14 with a significant p-value of 0, indicating the violation of local independence. Considering other model fit statistics discussed above, the acceptable values for indices such as M2, RMSEA, SRMSR, and

Table 2. Absolute Fit Indices for the G-DINA Model.

Statistic	mean[stats]	max[stats]	max[z.stats]	p	adj. p
Proportion Correct	.0007	.0021	.5204	.6028	1
Transformed Correlation	.0187	.1370	15.1081	.0000	0
Log Odds Ratio	.0869	.6339	13.9353	.0000	0

PC at least partially support the use of the model despite the violation of local independence. However, as warned by Santos and de la Torre (2019), the presence of local dependence results in biased parameter estimates. Thus, the results should be interpreted with caution.

4.2. DIF Analysis Results

G-DINA, as the model providing the best fit, was used to perform DIF analysis based on the Wald test and LRT. Appendix E displays the test statistics for the Wald test and LRT and the effect size measures for the flagged items. The effect size measure introduced by George and Robitzsch (2014) is presented under the "ABS d" column, and the one proposed by the authors to be used within the CDM-DIF context is shown in the column titled "d^2." The p-values were adjusted using Bonferroni correction to control for the Type I error due to multiple comparisons.

As shown in Appendix E, both methods flagged a large number of items. The results of the Wald test indicated that 8 out of 15 items displayed DIF at the .05 significance level. LRT flagged the same items as the Wald test except for Item 12, which had a nonsignificant p-value. Taking the effect size into account, Item 13 displayed negligible DIF, items of which type can be freely included in tests according to the ETS classification (Wiberg, 2007). Overall, the absolute and the squared difference effect size measures yielded almost similar results. However, Items 3, 9, and 12 were not consistently categorized by the two measures. Items 3 and 9 were classified as having medium DIF by the effect size based on the absolute difference, but displayed large DIF according to the squared distance effect size measure. In either event, ETS warns against using items showing medium and large DIF (Zieky, 1993). Item 12, which was only flagged by the Wald test, displayed negligible and moderate DIF based on the absolute and squared difference effect size measures, respectively.

Chi-square test statistics, LRT included, are notorious for being highly sensitive to sample size. To account for the large sample size that leads to a higher rejection rate for the null hypothesis of "no DIF" for the items, LRT statistics were adjusted for sample sizes equal to half, a quarter, and one-sixth of the original sample size to check the consistency of the results. The p-values were also adjusted accordingly, as presented in Table 3. Adjusting for the sample size resulted in fewer DIF items. Halving the sample size, Items 1, 6, and 8, which displayed large DIF in both effect size measures, had significant p-values. When adjusting the p-values for $N/4$ and $N/6$, only Item 1 displayed DIF.

Table 3. LRT Statistics Adjusted for Sample Size.

N		*N*/2 = 6084			*N*/4 = 3042			*N*/6 = 2028		
Item	LRStat	df	*p*	adj. *p*	LRStat	*p*	adj. *p*	LRStat	*p*	adj. *p*
1	58.08	4	.00	.00*	29.04	.00	.00*	19.36	.00	.01*
2	3.86	8	.86	1.00	1.93	.98	1.00	1.28	.99	1.00
3	18.61	8	.01	.25	9.30	.31	1.00	6.20	.62	1.00
4	17.32	8	.02	.40	8.66	.37	1.00	5.77	.67	1.00
5	6.51	8	.58	1.00	3.25	.91	1.00	2.17	.97	1.00
6	20.57	4	.00	.00*	10.28	.03	.53	6.85	.14	1.00
7	8.69	8	.36	1.00	4.34	.82	1.00	2.89	.94	1.00
8	28.35	8	.00	.00*	14.17	.07	1.00	9.45	.30	1.00
9	21.83	8	.00	.07	10.91	.20	1.00	7.27	.50	1.00
10	7.64	4	.10	1.00	3.82	.43	1.00	2.54	.63	1.00
11	0.57	2	.74	1.00	0.28	.86	1.00	0.19	.90	1.00
12	7.24	4	.12	1.00	3.62	.45	1.00	2.41	.65	1.00
13	12.50	4	.01	.20	6.25	.18	1.00	4.16	.38	1.00
14	5.30	2	.07	1.00	2.65	.26	1.00	1.76	.41	1.00
15	2.85	4	.58	1.00	1.42	.83	1.00	0.95	.91	1.00

Note. Flagged items are shown with the star symbol.

Table 4. Attribute Prevalence Estimates for the Comparison Groups.

Attribute	Female	Male
Vocabulary	.33	.42
Syntax	.38	.43
Main Idea	.28	.34
Inference	.33	.43
Detail	.34	.38

The attributes measured by DIF items were further inspected to gain a better understanding of those items. To do so, the researchers only focused on Items 1, 6, and 8 – the three items having large effect sizes, as discussing all seven items would not be feasible given the limited space. Vocabulary was the attribute that was required by all three items. Items 1 and 6 measured Vocabulary and Main Idea, whereas Item 8 measured Vocabulary, Syntax, and Detail.

For each item, bar charts of success probabilities by gender (Appendix F) were examined for latent classes with non-overlapping standard errors. For Item 8, the latent group "100" was not displayed in this figure due to having a less than one expected count, which resulted in unreliable estimates. As evident from the bar charts, females had uniformly lower probabilities of success across all mastery profiles.

Attribute prevalence for males and females was checked next to find out more about the attributes that were easier to master for the two genders. As shown in Table 4, Inference and Syntax were the easiest attributes for males as they were mastered by 43% of the examinees. The easiest attribute for females was Syntax which was mastered by 38% of the examinees. On the other hand, the most difficult attribute for both genders was Main Idea, which was mastered by 34% and 28% of the males and females, respectively. Females had consistently lower mastery rates in comparison to males across all the five attributes.

Prevalent latent classes were also examined to determine which attribute mastery profiles were the most common across the two genders. As presented in Appendix G, 36% of the examinees were nonmasters of all the attributes (latent class "00000") among males, and 19% of them had mastered all the attributes (latent class "11111"). The third most prevalent latent class was "11000" meaning that 12% of the examinees had mastered Vocabulary and Syntax. The same most prevalent attribute mastery profiles were observed among females. However, compared to males, a larger proportion of females belonged to the latent class "00000" (41%), and fewer females had mastered all the five attributes (13%).

4.3. Improvement in the Model Fit

Following the identification of DIF items, we checked whether or not the model that allowed the three DIF items' parameters to vary across males and females would fit the data significantly better than the original calibration of the data under single-group analysis. To this end, relative fit statistics for the four nested models were compared, wherein the most complex one was the MG G-DINA that assumed non-invariance of item parameters and allowed for separate estimation of item parameters for the two studied groups. The second MG model assumed invariant item parameters for the twelve non-DIF items and allowed separate parameter estimates for the three DIF items. The third model was an MG model that assumed item invariance for all the items across the two genders, and the fourth one, a single-group model, assumed the respondents did not belong to different manifest groups and thus, estimated one set of item parameters. Logically, it is expected that a multiple-group model allowing the parameters to vary across males and females provides the optimal fit. The likelihood ratio test was used to determine whether a simpler model could replace the MG G-DINA that estimated all the item parameters for the two genders separately.

As presented in Table 5, the MG model with the non-invariance assumption yielded the largest log-likelihood, followed by the less complex models with fewer numbers of parameters. The χ^2 values associated with the LRT all had significant p-values meaning that the simpler models could not

Table 5. Relative Fit Indices Comparing the Four Nested Models.

Model	# par	−2LL	AIC	BIC	SABIC	χ^2	*df*	*p*
MG-Separate	230	193,636	194,096	195,799	195,068			
12Common + 3DIF	162	193,765	194,089	195,280	194,774	129	68	<.001
MG-Common	146	193,879	194,171	195,253	194,789	115	16	<.001
Single-Group	115	194,343	194,572	195,424	195,058	464	31	<.001

provide an equally good fit to the model. Despite the significant *p*-values, the model "12Common + 3DIF" yielded smaller AIC and SABIC values, providing partial support for this model. For the same reason previously discussed concerning the sensitivity of χ^2-based tests to the sample size, the values of the χ^2 and log-likelihood were adjusted for the large sample size (not shown). Specifically, by adjusting the statistics for half the sample size, the χ^2 statistic for the LRT yielded a nonsignificant *p*-value for the model that allowed the item parameters for the three DIF items to be separately estimated for the two genders. This meant that the "12Common + 3DIF" model could fit the data equally well and replace the "MG-Separate" model.

5. Discussion

The present study inspected gender DIF in a high stakes RC test using CDMs. DIF analysis is a routine procedure for ascertaining fairness and establishing the validity of score interpretations. To this end, the attributes measured by the test were defined, and the initial Q-matrix was developed, validated, and used to select the model that best fitted the data – G-DINA. It should be noted that the utilized RC test was not developed for diagnostic purposes; therefore, we retrofitted G-DINA to it (Ravand & Baghaei, 2019). Cognitive diagnostic assessments, designed for diagnostic purposes from scratch, provide more detailed information about the examinees' performance. Retrofitting CDMs to already developed tests can still provide finer-grained information compared to single ability scores reported as in IRT analyses. Despite having a similar θ score, the students might belong to different attribute mastery profiles or students having the same profile might have distinct θ scores (Ma, Minchen, & de la Torre, 2020a). However, to benefit from such information, one must ensure the attributes are measured by sufficient numbers of items and that the model fits the data well. Hence, the fit of G-DINA was assessed in comparison to DINA, DINO, RRUM, LLM, and ACDM in the present study.

Fitting the MG G-DINA model, DIF was analyzed using the Wald test and LRT, and the effect sizes for the flagged items were calculated. The Wald test and LRT identified eight and seven DIF items, respectively, among which three items displayed large DIF according to both effect sizes. Adjusting LRT test statistics for half the sample size, the same three items (i.e., items 1, 6, & 8) had significant *p*-values. We utilized the CDM framework to delve into DIF items to obtain fine-grained information about the items' cognitive demands and the examinees' performances, and the cognitive processes involved in answering an item. CDM lends a fresh perspective on the reasons for the occurrence of DIF and possible explanations about the differential performance of the examinees. Thus, the three flagged items were further examined in detail to identify the attributes they measured. The three DIF items measured four out of the five attributes measured by the test. As the next step, success probabilities by gender were inspected at every latent class, and males were found to outperform females across all latent classes, which meant uniform DIF in favor of males for the three items.

The majority of DIF analyses have reported that, conditional on the same ability level, RC tests generally favor either females or none of the genders, which is at odds with the findings of the current study wherein DIF items favored males. The above discrepancy can be addressed from several perspectives. First, despite the fact that RC items have been found to generally favor females, finding items favoring males is not unprecedented. Lin and Wu (2004) who analyzed the English Proficiency Test in China using SIBTEST found three items that favored females, in contrast to the one item favoring males. Analyzing the verbal reasoning and English sections of the Psychometric Entrance Test in Israel, Gafni (1991) reported that the two studied forms displayed DIF in favor of males.

Findings of a study carried out by Ravand, Rohani, and Firoozi (2019) using multiple-indicators multiple-causes structural equation modeling showed that RC items favored males in an exam administered to select examinees into master's programs. Moreover, Bordbar (2020) examined the IUEE using Rasch model and found three items in the RC section favoring males, in contrast to the two items that favored females. Bordbar (2020) also examined the RC part of the general English section of the same test and found four items functioning differentially in favor of males. The above findings provide evidence that RC items can favor males.

Second, the present authors examined the overall performance of males and females based on the PC scores. Results showed that males outperformed females as evident from their PC scores, 0.37 vs. 0.29, which rules out the possibility of the CDM results' being merely an artifact of the method used. Third, previous studies on differential performance of males and females on RC tests have investigated the effect of the content of the passage, as well as the item on the performance of the two genders (Brantmeier, 2001; Bügel & Buunk, 1996). Analyzed in the light of schema theory, it is believed that prior knowledge of a topic facilitates processing of the new information (Rumelhart, 1980). Such interpretations are, however, post hoc explanations provided by the authors rather than content experts and are, thus, subjective. Therefore, care should be taken when interpreting the results. Previous studies have extensively reported that females perform better on items related to the mood, impression, or tone of the passage (Pae, 2012), human relations (O'Neill & McPeek, 1993), and passages on humanities (Doolittle & Welch, 1989). Males, on the other hand, have been reported to outperform females on items related to science topics and technical contents, and items requiring logical inference (Doolittle & Welch, 1989; Gafni, 1991; Lawrence, Curley, & McHale, 1988).

Content analysis of the DIF items was done by the present authors to possibly explain the reasons for the occurrence of DIF. In this study, the two passages whose items displayed DIF were about the stages of the expansion of cities and the Apollo moon landing. In particular, Items 6 and 8 were related to the second passage, which was on a science-related topic. This makes the items potentially easier for males. To be more specific, Item 6 asked for the main idea of the passage, which requires a general understanding of the whole text. Afflerbach's (1990) study showed that topic familiarity positively influenced the reformulation of the main idea of the passages. Thus, differential functioning of the item can be attributed to the fact that the topic of the passage is more male-friendly. Differential performance of Item 8 in favor of males can also be explained given that it requires skimming through the same science-related passage.

Item 1 based on the first passage tested the examinees' main idea comprehension on a topic related to urban development. As the item did not touch a content more familiar to males, the present authors could not come up with an explanation for differential functioning of this item. However, the fact that Items 1 and 6 both required the same skills of Vocabulary and Main Idea may shed light on this issue. Aside from topic familiarity, vocabulary plays a central role in understanding a passage and answering its items. Both passages as well as the four options in Items 6 and 8 were replete with formal words that could hinder understanding. Previous studies such as Edelenbos and Vinjé (2000) have reported that males perform better than females in vocabulary knowledge. However, the role of required skills in answering RC items and in the occurrence of DIF is an issue future studies can investigate in the light of CDMs.

We also examined attribute prevalence across genders and found that males had consistently higher attribute mastery rates across the five attributes. Inference and Syntax were the easiest attributes for males, and Syntax was the easiest skill for females to master. Main Idea was the most difficult skill to master for both genders. Detail and Vocabulary were the second most difficult attributes for males and females, respectively. Based on Bloom's taxonomy of comprehension levels (Bloom, 1956), recalling Detail and Main Idea requires lower-order cognitive processes, and thus, these two skills are easier to master than Inference. The findings of the present study are, therefore, not in line with Bloom's levels of cognition. The findings are worth considering more carefully as two out of the three DIF items measured only Vocabulary and Mail Idea. Detailed scrutiny of the reason for this discrepancy can be considered as potential future research.

Finally, to check whether taking into account the three DIF items improves model fit, three nested MG G-DINA models were compared with the single-group G-DINA model. Using the entire sample, none of the models provided a better fit than the MG G-DINA model with separate parameter estimates. When the χ^2 statistics were adjusted for half the sample size, however, the model accounting for the three DIF items had a nonsignificant *p*-value indicating that it performed as well as the MG G-DINA with non-invariant item parameters. Finding DIF items is the ultimate goal in many studies. However, allowing DIF analysis results to inform model selection would enable us to minimize the effect of DIF on item and person parameter estimates.

Findings of the present study inform the practice of test development in a number of ways. First, most tests, IUEE included, undergo a sensitivity review by a panel of experts, the aim of which is to identify potential sources of bias against a particular group of test takers. However, often, the sensitivity review panel experts cannot identify all sources of bias. Hence, the need to carry out statistical DIF analysis. Identified DIF items can be further examined by content experts and if found to be testing unintended constructs, the items can be revised or discarded from the test (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014). Second, CDM DIF analysis provides researchers a more detailed picture of the items that function differentially; thus, enabling researchers to focus on the passage, the content of DIF items, and the attributes measured by them to "better understand item and group properties that are responsible for differential test performance, and consequently may lead to tests that are more fair," (Penfield & Camilli, 2007, p. 155) and more accurately reflect test takers' abilities.

Third, CDMs' aim of providing finer-grained diagnostic feedback to examinees is accomplished by estimating the attribute profile each examinee belongs to as well as classifying examinees into masters or non-masters of a particular attribute. Apart from threatening the fairness and construct validity of tests, the existence of even moderate levels of DIF seriously jeopardizes profile-level classification accuracy and can endanger attribute-level accuracy in case the magnitude and number of DIF is large (Paulsen, Svetina, Feng, & Valdivia, 2020). As a measure of reliability for CDAs (Cui, Gierl, & Chang, 2012), classification accuracy can be improved by avoiding the inclusion of DIF items which, in turn, results in more accurate feedback to test takers. Fourth, since MG G-DINA can accommodate separate estimation of item parameters for all or certain items, identifying DIF items helps having improved model-data fit, leading to more precise classification accuracy (Gao, 2014).

Finally, the attribute-level accuracy indices provided by CDMs make it possible to choose the best test that measures the intended attributes the most accurately (Cui et al., 2012). However, the attributes can be measured more reliably if the attributes and the relationship among them is decided upon in advance and CDAs are built based on the selected CDM. Since the present study retrofitted the G-DINA model to IUEE, the two attributes of Reference and Rhetorical Purpose had to be combined with Inference and Main Idea so that they were measured by enough number of items. If considered important attributes worth being measured, CDAs can be designed whose items reliably measure Reference and Rhetorical Purpose, too. Thus, high quality items can be developed from the scratch in a way that they preferably measure each attribute in isolation to gain higher levels of classification accuracy, or in combination with other attributes by enough test items (Liu, Huggins-Manley, & Bradshaw, 2017).

Future research may focus on the differences between males and females in the cognitive processes they use to answer RC items. Qualitative research may address each gender's cognitive processes and the attributes utilized to answer RC test items. MG CDMs can accommodate different Q-matrices for the studied groups if the two genders are found to seek recourse to different cognitive processes and attributes. One more line of research relates to examining differential attribute functioning (Milewski & Baron, 2002) to check whether examinees with similar ability levels have statistically different probabilities of mastering each attribute. Findings of the present study indicated that females generally had lower attribute mastery rates than males, and a greater portion of the females had mastered none of the attributes. It is worth researching whether or not the attributes function differentially since feedback obtained from differential attribute functioning analysis can help language educators focus

on developing attributes in which a particular group has a lower probability of success. Furthermore, the Wald test and LRT used to detect DIF and compare nested models are χ^2-based tests highly sensitive to the sample size, and the smallest differences become statistically significant. The effect of the sample size on the estimates can be investigated, too.

In sum, the present study aimed to explore DIF in an RC test and look into the items displaying DIF more carefully by inspecting probabilities of success at every latent class and examining the attributes DIF items measure. The utilized RC test was a high stakes test screening examinees who aim to enter undergraduate programs. DIF analysis helps ensure the fairness of the test and the validity of the decisions based on test results. Despite widespread applications of DIF analysis, little is known about the nature of DIF and the reasons for its occurrence (Svetina et al., 2017). Using the CDM DIF framework, we aimed to probe item demands and the cognitive processes involved in answering test items to shed light on the findings of DIF analysis. Furthermore, since the model fit is of utmost importance when retrofitting CDMs to already developed tests, CDM DIF analysis helps identify items whose parameters need to be estimated separately in the model, which may result in significant improvement in fit indices and, in turn, in item and person parameter estimates.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

ORCID

Roghayeh Mehrazmay http://orcid.org/0000-0002-8599-4942

References

Ackerman, T. A. (1987, April). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local Independence*. Paper presented at the Annual Meeting of the American Educational Research Association Washington, DC, United States.

Afflerbach, P. P. (1990). The influence of prior knowledge on expert readers' main idea construction strategies. *Reading Research Quarterly*, *25*(1), 31–46. doi:10.2307/747986.

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

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Birenbaum, M., Kelly, A. E., & Tatsuoka, K. (1993). Diagnosing knowledge states in algebra using the rule-space model. *Journal for Research in Mathematics Education*, *24*(5), 442–459. doi:10.1002/j.2333-8504.1992.tb01488.x.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals*. New York, NY: David McKay.

Bordbar, S. (2020). Gender differential item functioning (GDIF) analysis in iran's university entrance exam. *English Language in Focus (ELIF)*, *3*(1), 49–68. doi:10.24853/elif.3.1.49-68.

Brantmeier, C. (2001). Second language reading research on passage content and gender: Challenges for the intermediate-level curriculum. *Foreign Language Annals*, *34*(4), 325–333. doi:10.1111/j.1944-9720.2001.tb02064.x.

Bügel, K., & Buunk, B. P. (1996). Sex differences in foreign language text comprehension: The role of interests and prior knowledge. *The Modern Language Journal*, *80*(1), 15–30. doi:10.1111/j.1540-4781.1996.tb01133.x.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, *37*(6), 419–437. doi:10.1177/0146621613479818.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*(2), 123–140. doi:10.1111/j.1745-3984.2012.00185.x.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*(8), 598–618. doi:10.1177/0146621613488436.

Cohen, A. D., & Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks* (TOEFL Monograph No. MS-33). Princeton, NJ: Educational Testing Service.

Cui, Y., Gierl, M., & Chang, H. –. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*(1), 19–38. doi:10.1111/j.1745-3984.2011.00158.x.

Day, R. R., & Park, J.-S. (2005). Developing reading comprehension questions. *Reading in a Foreign Language*, *17*(1), 60–73.

De Ayala, R. J. (2008). *The theory and practice of item response theory*. New York, NY: Guilford Press.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199. doi:10.1007/s11336-011-9207-7.

de la Torre, J., & Akbay, L. (2019). Implementation of cognitive diagnosis modeling using the GDINA R package. *Eurasian Journal of Educational Research*, *80*, 171–192. doi:10.14689/ejer.2019.80.9.

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*(2), 253–273. doi:10.1007/s11336-015-9467-8.

de la Torre, J., & Lee, Y. S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement*, *47*(1), 115–127. doi:10.1111/j.1745-3984.2009.00102.x.

DiBello, L., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics* (pp. 979–1030). Amsterdam, the Netherlands: Elsevier.

Doolittle, A., & Welch, C. (1989, March). *Gender differences in performance on a college-level achievement test*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, United States.

Edelenbos, P., & Vinjé, M. (2000). The assessment of a foreign language at the end of primary (elementary) education. *Language Testing*, *17*(2), 144–162. doi:10.1177/026553220001700203.

Gafni, N. (1991, April). *Differential item functioning: Performance by sex on reading comprehension tests*. Paper presented at the Annual Meeting of the Academic Committee for Research on Language Testing, Kiryat Anavim, Israel.

Gao, M. (2014). *Assessing the model fit and classification accuracy in cognitive diagnosis models* (Unpublished doctoral dissertation). University of California, Gainesville, FL.

George, A. C., & Robitzsch, A. (2014). Multiple group cognitive diagnosis models, with an emphasis on differential item functioning. *Psychological Test and Assessment Modeling*, *56*(4), 405–432. doi:10.1177/0146621620965745.

Gierl, M. J., & Leighton, J. P. (2007). Directions for future research in cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 341–351). Cambridge: Cambridge University Press.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 225–252. doi:10.1111/j.2044-8317.2012.02050.x.

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

Henson, R. A., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210. doi:10.1007/s11336-008-9089-5.

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*(1), 98–125. doi:10.1111/jedm.12036.

Jang, E., Dunlop, M., Park, G., & van der Boomet, E. H. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Language Testing*, *32*(3), 359–383. doi:10.1177/0265532215570924.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion model application to LanguEdge assessment. *Language Testing*, *26*(1), 31–73. doi:10.1177/0265532208097336.

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

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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. doi:10.2307/2529310.

Langer, M. M. (2008). *A reexamination of Lord's Wald test for differential item Functioning using item response theory and modern error Estimation* (Unpublished doctoral dissertation), The University of North Carolina, Chapel Hill.

Lawrence, I. M., Curley, W. E., & McHale, F. J. (1988). *Differential item functioning for males and females on SAT verbal reading subscore items* (ETS Research Report No. RR-88-10). Princeton, NJ: Educational Testing Service.

Lee, Y. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing*, *21*(1), 74–100. doi:10.1191/0265532204lt260oa.

Lee, Y., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly*, *6*(3), 172–189. doi:10.1080/15434300902985108.

Lee, Y.-S., de la Torre, J., & Park, Y. S. (2012). Relationships between cognitive diagnosis, CTT, and IRT indices: An empirical investigation. *Asia Pacific Educational Review*, *13*(2), 333–345. doi:10.1007/s12564-011-9196-3.

Leighton, J. P., & Gierl, M. J. (2007). Why cognitive diagnostic assessment? In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 3–18). Cambridge: Cambridge University Press.

Li, F. (2008). *A modified higher-order DINA model for detecting differential item functioning and differential attribute functioning* (Unpublished doctoral dissertation), The University of Georgia, Athens.

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. doi:10.1177/0265532215590848.

Li, H., & Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment*, *18*(1), 1–23. doi:10.1080/10627197.2013.761522.

Li, X., & Wang, W.-C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement*, *52*(1), 28–54. doi:10.1111/jedm.12061.

Lin, J., & Wu, F. (2004, March). *Differential performance by gender in foreign language Testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL, United States.

Lissitz, B., Jiao, H., Li, M., Lee, D. Y., & Kang, Y. (2014). Cognitive diagnostic models: Executive report for the maryland state department of education by the MARC team. Retrieved from https://marces.org/current/ExecutiveReport_MARC_2014_Cognitive%20Diagnostic%20Models.pdf .

Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2017). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and Psychological Measurement*, *77*(2), 220–240. doi:10.1177/0013164416645636.

Ma, W., & de la Torre, J. (2020). GDINA: The generalized DINA model framework. R package version 2.8.0. Retrieved from https://CRAN.R-project.org/package=GDINA

Ma, W., Minchen, N., & de la Torre, J. (2020a). Choosing between CDM and unidimensional IRT: The proportional reasoning test case. *Measurement: Interdisciplinary Research and Perspectives*, *18*(2), 87–96. doi:10.1080/15366367.2019.1697122.

Ma, W., Terzi, R., & de la Torre, J. (2020b). Detecting differential item functioning using multiple-group cognitive diagnosis models. *Applied Psychological Measurement*, *45*(1), 37–53. doi:10.1177/0146621620965745.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719–748. doi:10.1093/jnci/22.4.719.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212. doi:10.1007/bf02294535.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 71–101. doi:10.1080/15366367.2013.831680.

Milewski, G. B., & Baron, P. A. (2002, April). *Extending DIF methods to inform aggregate reports on cognitive skills*. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA, United States.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Nájera, P., Sorrel, M., & Abad, P. (2019). Reconsidering cutoff points in the general method of empirical q-matrix validation. *Educational and Psychological Measurement*, *79*(4), 727–753. doi:10.1177/0013164418822700.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Oshima, T. C., & Morris, S. B. (2008). Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice*, *27*(3), 43–50. doi:10.1111/j.1745-3992.2008.00127.x.

Pae, T.-I. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing*, *29*(4), 533–554. doi:10.1177/0265532211434027.

Paulsen, J., Svetina, D., Feng, Y., & Valdivia, M. (2020). Examining the impact of differential item functioning on classification accuracy in cognitive diagnostic models. *Applied Psychological Measurement*, *44*(4), 267–281. doi:10.1177/0146621619858675.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C. R. Rao (Eds.), *Handbook of statistics (Vol. 26): Psychometrics* (pp. 125–167). New York, NY: Elsevier.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197–207. doi:10.1177/014662169001400208.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, *34*(8), 782–799. doi: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. doi:10.1080/15305058.2019.1588278.

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., Rohani, G., & Firoozi, T. (2019). Investigating gender and major DIF in the Iranian national university entrance exam using multiple-indicators multiple-causes structural equation modelling. *Issues in Language Teaching*, *8*(1), 33–61. doi:10.22054/ilt.2020.49509.460.

Robitzsch, A., Kiefer, T., George, A. C., & Ünlü, A. (2020). CDM: Cognitive diagnosis modeling. R package version 7.5-15. Retrieved from https://CRAN.R-project.org/package=CDM

Roussos, L. A., DiBello, L. V., Henson, R. A., Jang, E., & Templin, J. L. (2010). Skills diagnosis for education and psychology with IRT-based parametric latent class models. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 35–69). Washington, USA: American Psychological Association.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33–58). Hillsdale, NJ: Erlbaum.

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

Santos, K. C., & de la Torre, J. (2019, April). *The impact of conditional dependency and its detection in cognitive diagnosis modeling*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Toronto, Canada.

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*(3), 190–209. doi:10.1080/15434300903559209.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education*, *2*(3), 255–275. doi:10.1207/s15324818ame0203_6.

Shealy, R. T., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194. doi:10.1007/bf02294572.

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *19*(2–3), 170–187. doi:10.1080/13803611.2013.767621.

Stout, W. F., Nandakumar, R., & Habing, B. (1996). Analysis of latent dimensionality of dichotomously and polytomously scored test data. *Behaviormetrika*, *23*(1), 37–65. doi:10.2333/bhmk.23.37.

Svetina, D., Dai, S., & Wang, X. (2017). Use of cognitive diagnostic model to study differential item functioning in accommodations. *Behaviormetrika*, *44*(2), 313–349. doi:10.1007/s41237-017-0021-0.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354. doi:10.1111/j.1745-3984.1983.tb00212.x.

Tatsuoka, K. K. (1986, July). *Toward an integration of item response theory and cognitive error diagnoses*. Paper presented at the Educational Testing Service Conference, Princeton, NJ, United States.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. London, UK: Routledge.

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

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

Wang, Z., Guo, L., & Bian, Y. (2015). The impact of DIF on estimating accuracy of cognitive diagnostic test. *Psychological Exploration*, *35*(3), 1923–1932.

Wiberg, M. (2007). *Measuring and detecting differential item functioning in criterion- referenced licensing test: A theoretic comparison of methods* (EM No 60). Umeå: Department of educational measurement, Umeå universitet. Retrieved from http://coshima.davidrjfikis.com/EPRS9360/Articles/Wiberg_07.pdf

Wright, K. D., & Oshima, T. C. (2015). An effect size measure for Raju's differential functioning for items and tests. *Educational and Psychological Measurement*, *75*(2), 338–358. doi:10.1177/0013164414532944.

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

Zhang, W. (2006). *Detecting differential item functioning using the DINA model* (Unpublished doctoral dissertation). The University of North Carolina, Greensboro.

Zheng, C., & Chang, -H.-H.(2016). High-efficiency response distribution-based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, *40*(8), 608–624. doi:10.1177/0146621616665196.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (Ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Appendix A

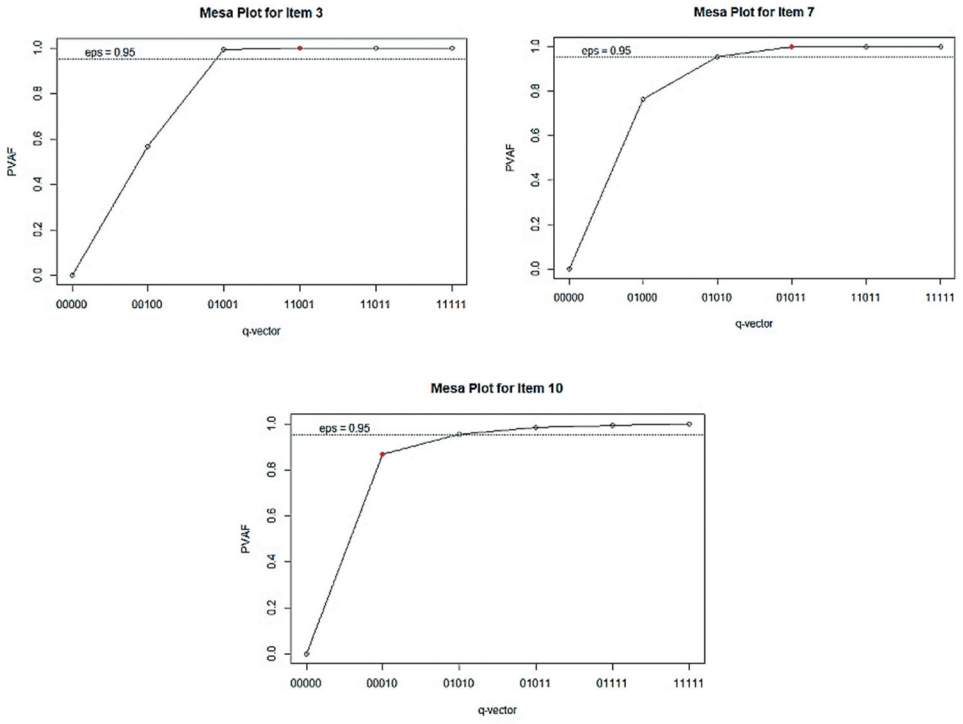
The Initial Q-matrix Based on the Experts' Coding.

		Attributes				
		Vocabulary	Syntax	Main Idea	Inference	Detail
Items	1	1*	0	1	0	0
	2	1	0	1	1	0
	3	0*	1	0	0	1
	4	0	1	0	1	1
	5	0	1*	1	0	1*
	6	1	0	1	0	0
	7	0	1	0	1	0*
	8	1	1	0	0	1
	9	0	1	0	1	1
	10	0	1*	0	1	0
	11	0	0	1	0	0
	12	0	0	0	1	1
	13	1	0	1	0	0
	14	0*	0	0	1	0
	15	1	0	0	0	1
	κ	.568	.545	.635	.679	.628

Note. The star symbols denote suggested modifications in the first run.

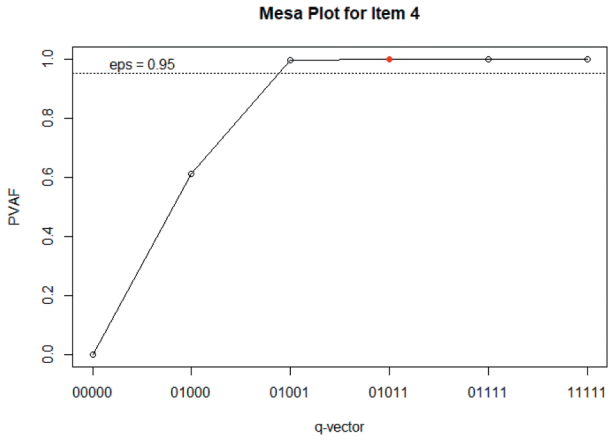
Appendix B

Mesa Plots for Items 3, 7, & 10 in the First Round of Q-Matrix Validation.



Appendix C

Mesa Plot for Items 4 in the Second Round of Q-Matrix Validation.



Appendix D

Relative Fit Indices Comparing G-DINA to Other Reduced Models.

Model	#par	logLik	Deviance	AIC	BIC	χ^2	df	*p*
G-DINA	115	−97,171.06	194,342.13	194,572.13	195,423.89			
DINA	61	−98,363.55	196,727.09	196,849.09	197,300.90	2384.97	54	<.001
DINO	61	−98,490.12	196,980.25	197,102.25	197,554.06	2638.12	54	<.001
ACDM	81	−97,490.51	194,981.03	195,143.03	195,742.96	638.9	34	<.001
RRUM	81	−97,423.09	194,846.19	195,008.19	195,608.13	504.06	34	<.001
LLM	81	−97,332.18	194,664.35	194,826.35	195,426.29	322.23	34	<.001

Appendix E

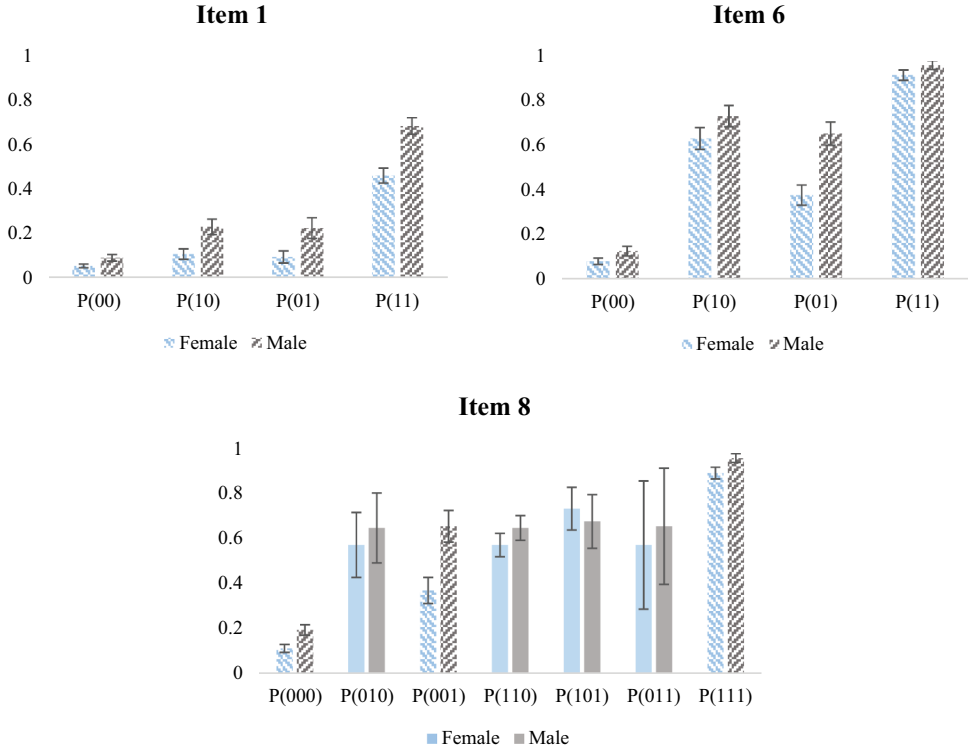
The Results of the Wald Test and LRT

	Wald			LRT			Effect Size			
Item	WaldStat.	*p*	adj. *p*	LRstat	*p*	adj. *p*	ABD *d*		d^2	
1	125.88	.00	.00*	116.16	.00	0.00*	.099	L	.013	L
2	7.62	.47	1.00	7.73	.46	1.00				
3	165,317.60	.00	.00*	37.24	.00	0.00*	.066	M	.008	L
4	36.48	.00	.00*	34.64	.00	0.00*	.066	M	.007	M
5	11.80	.16	1.00	13.04	.11	1.00				
6	54.60	.00	.00*	41.15	.00	0.00*	.089	L	.013	L
7	12.54	.13	1.00	17.40	.03	0.39				
8	386,011.82	.00	.00*	56.70	.00	0.00*	.104	L	.017	L
9	39.37	.00	.00*	43.66	.00	0.00*	.065	M	.008	L
10	13.19	.01	.16	15.29	.00	0.06				
11	1.33	.52	1.00	1.15	.56	1.00				
12	17.33	.00	.03*	14.49	.01	0.09	.053	N	.007	M
13	26.54	.00	.00*	25.00	.00	0.00*	.040	N	.002	N
14	8.32	.02	.23	10.60	.01	0.07				
15	6.35	.17	1.00	5.70	.22	1.00				

Note. Flagged items are shown with a star symbol.

Appendix F

Bar Charts of Success Probabilities per Gender and Latent Class.



Note. Solid bars represent latent classes whose standard errors overlap, and the striped bars represent latent classes with non-overlapping standard errors.

Appendix G

Attribute Mastery Profiles per Gender

Latent Class	Posterior Probability		Latent Class	Posterior Probability	
	Females	Males		Females	Males
00000	**.41**	**.36**	11100	.00	.00
10000	.00	.00	11010	.02	.04
01000	.03	.01	11001	.00	.00
00100	.00	.01	10110	.00	.00
00010	.01	.04	10101	.00	.00
00001	.06	.01	10011	.03	.04
11000	**.12**	**.12**	01110	.02	.02
10100	.00	.00	01101	.01	.01
10010	.00	.00	01011	.00	.00
10001	.00	.01	00111	.05	.06
01100	.00	.00	11110	.00	.00
01010	.02	.01	11101	.00	.00
01001	.00	.00	11011	.01	.01
00110	.01	.00	10111	.02	.00
00101	.05	.04	01111	.00	.01
00011	.00	.00	**11111**	**.13**	**.19**

Note. The three most prevalent latent classes are highlighted in bold type.

Appendix H

Codes Used in the Study

```
# Importing data and the Q-matrix
## The first column of the data file includes the gender of the cases.
dat <- read.csv ("E:/Data/data.csv," header = TRUE)
Q <- read.csv ("E:/Data/Q.csv," header = TRUE)
# Q-matrix validation using de la Torre and Chiu's method, Round I
fit <- GDINA (dat = dat[,-1], Q = Q, model = "GDINA")
qval1 <- Qval (fit, method = "PVAF," eps = 0.95)
PVAFvalues1 <- extract (qval1, what = "PVAF")
PVAFvalues1
suggQ1 <- extract (qval, what = "sug.Q")
suggQ1
# Draw mesa plots using the function plot
plot (qval1, item = 3)
plot (qval1, item = 7)
plot (qval1, item = 10)
# Q-matrix validation, Round II
NewQ <- read.csv ("E:/Data/NewQ.csv," header = T)
fit <- GDINA (dat = dat[,-1], Q = NewQ, model = "GDINA")
qval2 <- Qval (fit, method = "PVAF," eps = 0.95)
PVAFvalues2 <- extract (qval1, what = "PVAF")
PVAFvalues2
suggQ2 <- extract (qval1, what = "sug.Q")
suggQ2
# Draw mesa plots using the function plot
plot (qval2, item = 4)
# Model fit
fit <- GDINA (dat = dat[,-1], Q = NewQ, model = "GDINA," mono.constraint = TRUE)
fitDINA <- GDINA (dat = dat[,-1], Q = NewQ, model = "DINA," mono.constraint = TRUE)
fitDINO <- GDINA (dat = dat[,-1], Q = NewQ, model = "DINO," mono.constraint = TRUE)
```

```
fitACDM <- GDINA (dat = dat[,-1], Q = NewQ, model = "ACDM," mono.constraint = TRUE)
fitRRUM <- GDINA (dat = dat[,-1], Q = NewQ, model = "RRUM," mono.constraint = TRUE)
fitLLM <- GDINA (dat = dat[,-1], Q = NewQ, model = "LLM," mono.constraint = TRUE)
## Relative Fit indices
anova (fit, fitDINA, fitDINO, fitACDM, fitRRUM, fitLLM,)
## Absolute Fit indices
summary (fit)
# Local independence
LD <- CDM::gdina(dat = dat[,-1], q.matrix = NewQ, group = dat[,1], rule = "GDINA")
summary (LD)
summary (CDM::modelfit.cor.din (LD))
# DIF
wald <- dif (dat[,-1], Q = New Q, group = dat[,1], method = "wald," p.adjust.methods = "bonferroni," nstarts =100,
mono.constraint = TRUE, SE.type = 3)
wald
# DIF using LR test
LR <- dif (dat[,-1], Q = New Q, group = dat[,1], method = "LR," p.adjust.methods = "bonferroni," nstarts = 100,
mono.constraint = TRUE, SE.type = 3)
LR
# Improvement
## Fit indices for the "MG-Separate" model
### Create a block diagonal matrix of data and its corresponding Q-matrix, and import them
mgSep<- GDINA (dat = dat30[,-1], Q = QM30, group = dat30[,1], nstarts = 5)
## Fit indices for the "MG-Common" model
mgCom<- GDINA(dat = dat[,-1], Q = New Q, group = dat[,1], nstarts = 5)
## Fit indices for the "12 Common + 3 DIF" model
### Create a block diagonal matrix of data with three DIF items and 12 common items. Then, create its
corresponding Q-matrix, and import them.
DIF<- GDINA (dat = datDIF[,-1], Q = QDIF, group = datDIF[,1])
# Compare relative model fit indices
anova (fit, mgSep, mgCom, DIF)
```