International Journal of Assessment Tools in Education
2020, Vol. 7, No. 2, 145–158
https://doi.org/10.21449/ijate.689752

Published at https://dergipark.org.tr/en/pub/ijate *Research Article*

Wald Test Formulations in DIF Detection of CDM Data with the Proportional Reasoning Test

Likun Hou [1], **Ragip Terzi** [2,*], **Jimmy de la Torre** [3]

[1] Educational Testing Service, NJ, USA
[2] Department of Educational Measurement and Evaluation, Harran University, Sanliurfa, Turkey
[3] Division of Learning, Development and Diversity, The University of Hong Kong, Pokfulam, Hong Kong

ARTICLE HISTORY

Received: Feb 16, 2020
Revised: Apr 7, 2020
Accepted: Apr 17, 2020

KEYWORDS

Cognitive Diagnosis Model,
Proportional Reasoning,
Wald Test,
DIF

Abstract: This study aims to conduct differential item functioning analyses in the context of cognitive diagnosis assessments using various formulations of the Wald test. In implementing the Wald test, two scenarios are considered: one where the underlying reduced model can be assumed; and another where a saturated CDM is used. Illustration of the different Wald test to detect DIF in CDM data was based on the items' performance of the Proportional Reasoning test among low- and high-performing school students. A benchmark simulation study was included to compare the performance of the Wald test in each scenario. The agreement of the latent attribute classification based on different cognitive diagnosis models was also discussed.

1. INTRODUCTION

Cognitive diagnosis models (CDMs) are a family of multidimensional latent class models that are used to obtain finer grained information on students' learning progress. CDMs classify examinees based on attribute mastery profiles that determine students' membership in latent groups. Each latent group is denoted by a binary vector with 1s and 0s, indicating mastery and nonmastery of each of the attributes being measured, respectively.

To date, despite the benefits of cognitive diagnosis assessments (CDAs), the application of CDMs has been limited. Some researchers (Tatsuoka, 1984; Tjoe & de la Torre, 2014) have created some tests based on CDA through an intensive study. In these studies, specific latent attributes were constructed as finer-grained and interrelated, but separable skills within a domain of interest. However, many psychometric questions about the CDM framework still remain. One such question is about differential item functioning (DIF) in CDMs. DIF analyses are regularly carried out for the purpose of test fairness and validity (Camilli, 2006). In the context of CDMs, DIF occurs if students with the same attribute mastery profile but from distinct observed groups have different probabilities of correctly answering an item (Hou, de la

CONTACT: Ragip Terzi ✉ terziragip@harran.edu.tr 🖃 Harran University, School of Education, Sanliurfa, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

Torre, & Nandakumar, 2014; Li, 2008). DIF analysis is necessary to examine parameter or construct invariance (Zumbo, 2007). Invariance pertains to the item responses that should be independent conditioned on attribute profiles. Therefore, DIF analysis is important to investigate the invariance of attribute-item interactions across groups (Hou et al., 2014).

Currently, there exist a few studies for DIF detection purposes in CDMs (e.g., Hou et al., 2014; Li, 2008; Milewski & Baron, 2002; Zhang, 2006). Milewski and Baron (2002) examined group differences in skill mastery profiles controlling for overall ability where skill strengths and weaknesses were analyzed. However, they did not investigate whether an item was biased due to a specific skill. Furthermore, the Mantel-Haenszel (MH; Holland & Thayer, 1988) and SIBTEST methods (Shealy & Stout, 1993) were applied by Zhang (2006) to examine DIF for the deterministic inputs, noisy "and" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model based on total test scores and attribute profile scores. However, the two methods were limited to detect only uniform DIF. Moreover, the estimates of the item parameters and attribute mastery profiles were contaminated because of including potential DIF items in the procedures. The study of Milewski and Baron (2002) was extended by Li (2008) to a modified higher-order DINA model (HO-DINA; de la Torre & Douglas, 2004), where DIF and differential attribute functioning (DAF) were simultaneously investigated. In addition to DIF described previously, DAF occurs if students with the same attribute mastery profile but from different observed groups have different probabilities of mastering an attribute. The higher-order (HO) structure in this procedure explains the relationship among items, attributes, and general ability; however, it was also limited to uniform DIF detections. Given these limitations, Hou et al. (2014) introduced the Wald test for DIF detection purposes in the DINA model. This procedure has two major advantages. First, separate calibrations were performed for the reference (R) and focal (F) groups so as not to require test purification for DIF contaminations. Second, the procedure can effectively detect both uniform and nonuniform DIF. The Wald test also outperformed the MH and SIBTEST procedures in detecting uniform DIF.

This study aims to carry out DIF analyses in the context of CDMs using various formulations of the Wald test. In implementing the Wald test, two scenarios were considered: one where the underlying reduced model (i.e., DINA model) was assumed; another scenario where a saturated CDM was used. The purpose of this study is to illustrate the performance of the different Wald tests in detecting DIF in CDM data; thus, the Proportional Reasoning test data (Tjoe & de la Torre, 2014) for schools with different proficiency levels were used. In particular, DIF items are detected when the groups are defined as high-performing school versus low-performing school.

1.1. Theoretical Framework

1.1.1. *G-DINA Model*

In the last two decades, the DINA model has been a very commonly used reduced CDM. This model classifies examinees into two groups, those who do have and who do not have all the required attributes. In other words, missing any one of the required attributes is the same as missing all of them. However, this restriction may be too strict under certain situations. de la Torre (2011) proposed the generalized DINA (G-DINA) model where examinees are classified into 2^{K_j} latent groups, and K_j is the number of the required attributes for item j (i.e., $K_j = \sum_{k=1}^{K} q_{jk}$). Therefore, examinees who have mastered different attributes can have different probabilities of correctly answering an item.

Let item j require the first $1, \cdots, K_j$ attributes. The reduced attribute vector can be denoted by $\boldsymbol{\alpha}_{lj}^{*}$, which represents the columns of the required attributes (i.e., $l = 1, \cdots, 2^{K_j}$). $P(X_j = 1|\boldsymbol{\alpha}_{lj}^{*}) = P(\boldsymbol{\alpha}_{lj}^{*})$ can represent the probability of correctly answering an item j by examinees with attribute pattern $\boldsymbol{\alpha}^*_{lj}$. The item response function of the G-DINA model for the identity link is given by

$$P(\boldsymbol{\alpha}^*_{lj}) = \delta_{j0} + \sum_{k=1}^{K_j} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j} \sum_{k=1}^{K_j-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{j12\ldots K_j} \prod_{k=1}^{K_j} \alpha_{lk}, \quad (1)$$

where δ_{j0} is the intercept for item j; δ_{jk} is the main effect of α_k; $\delta_{jkk'}$ is the interaction effect of α_k and $\alpha_{k'}$; and $\delta_{j12\ldots K_j}$ is the interaction effect of $\alpha_1, \ldots, \alpha_{K_j}$.

The G-DINA model is a commonly used saturated model that subsumes several reduced CDMs such as the DINA model, the DINO model, the *A*-CDM, the LLM, and the R-RUM. These reduced models can be obtained from the G-DINA model by applying appropriate parameterization (de la Torre, 2011). For example, after setting all the parameters in Equation (1) to zero, except for δ_{j0} and $\delta_{j12\ldots K_j}$, the DINA model can be formulated as,

$$P(\boldsymbol{\alpha}^*_{lj}) = \delta_{j0} + \delta_{j12\ldots K_j} \prod_{k=1}^{K_j} \alpha_{lk}. \quad (2)$$

In this present study, the DINA and G-DINA models were employed as reduced and saturated models, respectively. The former model assumes a specific underlying process, whereas, the latter does not.

1.1.2. *The Wald Test*

The Wald test (Morrison, 1967) has been used in various statistical analyses for decades. In particular, the Wald test in the context of CDMs has been applied to a number of studies (de la Torre, 2011; de la Torre & Lee, 2013; Hou et al., 2014; Ma, Iaconangelo, & de la Torre, 2016; Terzi, 2017). The Wald test for CDM applications was first introduced by de la Torre (2011) to investigate whether the G-DINA model can be replaced by one of the reduced models (i.e., DINA, DINO, or *A*-CDM). The null hypothesis to test the fit of a reduced model with $p < 2^{K_j}$ parameters can be written as $\boldsymbol{R}_{jp} \times \boldsymbol{P}_j = 0$, where $\boldsymbol{P}_j = \{P(\boldsymbol{\alpha}^*_{lj})\}$, and $\boldsymbol{R}_{jp}$ is the $(2^{K_j} - p) \times 2^{K_j}$ restriction matrix. The Wald statistic W_j to test the null hypothesis for item j is computed as

$$W_j = [\boldsymbol{R}_{jp} \times \boldsymbol{P}_j]'[\boldsymbol{R}_{jp} \times Var(\boldsymbol{P}_j) \times \boldsymbol{R}'_{jp}]^{-1}[\boldsymbol{R}_{jp} \times \boldsymbol{P}_j], \quad (3)$$

where Var($\boldsymbol{P}_j$) is the variance-covariance matrix of the item parameters for the saturated model computed from the inverse of the information matrix. Under the null hypothesis for the DINA model (i.e., $p = 2$), the Wald statistic is assumed to be asymptotically χ^2 distributed with $2^{K_j} - p$ degrees of freedom.

Moreover, the Wald test has also been applied at the item level by comparing the fit of a saturated model to the fits of reduced models to come up with the most appropriate CDM (de la Torre & Lee, 2013). They found that the Wald test had excellent power to determine the true underlying model even for small sample sizes, while controlling the Type-I error for large sample sizes with a small number of attributes. The Wald test application in the study of de la Torre and Lee (2013) was extended by Ma et al. (2016), in that the Wald test was evaluated across several popular additive models and was shown that it can identify correct reduced models and improve attribute classifications. Hou et al. (2014) further carried out the Wald test for DIF detection in the context of CDMs, where the Wald test was able to detect both uniform and nonuniform DIF in the DINA model.

1.2. DIF in Cognitively Diagnostic Assessments

In contrast to IRT, DIF for CDMs needs to be redefined because the examinees are provided with the mastery profile of latent discrete attributes instead of locating examinees on the latent continuum. DIF in CDMs can be represented as $\triangle_{j\boldsymbol{\alpha}_l}= P(X_j = 1|\boldsymbol{\alpha}_l)_F - P(X_j = 1|\boldsymbol{\alpha}_l)_R$, where $\triangle_{j\boldsymbol{\alpha}_l}$ denotes DIF in item j for examinees with the attribute mastery profile $\boldsymbol{\alpha}_l$; $P(X_j = 1|\boldsymbol{\alpha}_l)_F$ is the success probability on item j for examinees with the attribute mastery profile $\boldsymbol{\alpha}_l$ in the F group; and similarly $P(X_j = 1|\boldsymbol{\alpha}_l)_R$ in the R group. There is no DIF if $\triangle_{j\boldsymbol{\alpha}_l}= 0$ for all attribute mastery profiles.

Because there are two parameters (the slip and guessing parameters) in the DINA model, DIF can be investigated by examining the differences in the slip and guessing parameters between the F and R groups. Item j exhibits DIF if:

$$\triangle_{s_j}= s_{R_j} - s_{F_j} \neq 0, \tag{4}$$

and/or

$$\triangle_{g_j}= g_{R_j} - g_{F_j} \neq 0. \tag{5}$$

For the G-DINA model, each item parameter corresponds to the probability of success on item j for examinees with the reduced attribute vector $\boldsymbol{\alpha}^*_{lj}$. Thus, DIF in the G-DINA model is the difference in the item parameters between the F and R groups, represented by $\triangle_{j\boldsymbol{\alpha}^*_{lj}}= P(X_j = 1|\boldsymbol{\alpha}^*_{lj})_F - P(X_j = 1|\boldsymbol{\alpha}^*_{lj})_R$, where $\triangle_{j\boldsymbol{\alpha}^*_{lj}}\neq 0$ denotes DIF in item j for examinees with the attribute mastery profile $\boldsymbol{\alpha}^*_{lj}$.

1.2.1. *The Wald Test for DIF Analysis*

The Wald test detects DIF in the CDM through multivariate hypothesis testing. To detect DIF in the DINA model, the null hypothesis is written as:

$$H_0\colon \begin{cases} s_{Fj} - s_{Rj} = 0 \\ g_{Fj} - g_{Rj} = 0^{\cdot} \end{cases} \tag{6}$$

The alternative hypothesis is that at least one of the item parameters is different between the F and R groups. There are two steps to implement the Wald test. In the first step, item parameters are calibrated for the F and R groups separately. The first step translates into applying an unconstrained model to the data, where no constraints in the item parameters across the F and R groups are used. The parameter estimates for item j across the two groups are represented as

$$\widehat{\boldsymbol{\beta}}^*_j = (\widehat{\boldsymbol{\beta}}_{Rj}, \widehat{\boldsymbol{\beta}}_{Fj}) = (\hat{g}_{Rj}, \hat{s}_{Rj}, \hat{g}_{Fj}, \hat{s}_{Fj})'. \tag{7}$$

In the second step, the null hypothesis of the equality of item parameters of the F and R groups is tested. The null hypothesis given in Equation (6) can be expressed in terms of the constrained model as follows:

$$H_0\colon \boldsymbol{R}_j \cdot \widehat{\boldsymbol{\beta}}^*_j = \boldsymbol{0}\,, \tag{8}$$

where $\boldsymbol{R}_j$ is a 2 × 4 matrix of restrictions, given as follows:

$$\boldsymbol{R}_j = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}. \tag{9}$$

The Wald statistic W_j to test the null hypothesis is computed as:

$$W_j = [\,\boldsymbol{R}_j \times \widehat{\boldsymbol{\beta}}^*_j]'[\boldsymbol{R}_j \times Var(\widehat{\boldsymbol{\beta}}^*_j) \times \boldsymbol{R}_j{}']^{-1}[\boldsymbol{R}_j \times \widehat{\boldsymbol{\beta}}^*_j], \tag{10}$$

where $Var(\widehat{\boldsymbol{\beta}}^*_j)$ is the variance-covariance matrix of the item parameters, written as:

$$Var(\widehat{\boldsymbol{\beta}}^*_j) = \begin{pmatrix} Var(\widehat{\boldsymbol{\beta}}_{Rj}) & 0 \\ 0 & Var(\widehat{\boldsymbol{\beta}}_{Fj}) \end{pmatrix}, \tag{11}$$

and under the null hypothesis $H_0\colon \boldsymbol{R}_j \cdot \boldsymbol{\beta}^*_j = \boldsymbol{0}$, and W_j is asymptotically χ^2 distributed with two degrees of freedom under the DINA model.

Similarly, in the G-DINA model, the first step of the Wald test is to estimate the item parameters separately for the F and R groups in the form of the vector written as follows:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^*_j &= (\widehat{\boldsymbol{\beta}}_{Rj}, \widehat{\boldsymbol{\beta}}_{Fj})' \\ &= (\hat{P}(\alpha^*_{\mathbf{0}j})_R, \cdots, \hat{P}(\alpha^*_{\boldsymbol{l}j})_R, \cdots, \hat{P}(\alpha^*_{\mathbf{1}j})_R, \hat{P}(\alpha^*_{\mathbf{0}j})_F, \cdots, \hat{P}(\alpha^*_{\boldsymbol{l}j})_F, \cdots, \hat{P}\big(\alpha^*_{\mathbf{1}j})_F\big)'. \end{aligned} \tag{12}$$

In the second step, the null hypothesis of the equality of item parameters of the F and R groups is tested, as in $H_0\colon \boldsymbol{R}_j \cdot \boldsymbol{\beta}^*_j = \boldsymbol{0}$. Since there are 2^{K_j} parameters to be estimated for each group, there are 2^{K_j} constraints and the dimension of the restriction matrix $\boldsymbol{R}_j$ is $2^{K_j} \times 2^{K_j+1}$. For example, for an item requiring two attributes for a correct response ($K_j = 2$), $\boldsymbol{R}_j$ is given as:

$$\boldsymbol{R}_j = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}. \tag{13}$$

Under $H_0\colon \boldsymbol{R}_j \cdot \boldsymbol{\beta}^*_j = \boldsymbol{0}$, the Wald statistic W_j in this example is assumed to be asymptotically χ^2 distributed with four degrees of freedom. Similar to the use of the Wald test for DIF detection in the DINA model, it only requires the estimation of the unconstrained model, that is, the item parameters are calibrated for the F and R groups separately.

In the G-DINA model, the Wald test can also be used to detect DIF when the underlying restricted model is specified (e.g. DINA model). It is carried out the same way as it is in the G-DINA model, but the restriction matrix $\boldsymbol{R}_j$ is structured differently, depending on which restricted model is assumed. For example, when the DINA model is assumed as the underlying restricted model, there are $2^{K_j+1} - 2$ constraints and the dimension of the restriction matrix $\boldsymbol{R}_j$ is $(2^{K_j+1} - 2) \times 2^{K_j+1}$. For an item requiring two attributes for a correct response ($K_j = 2$), $\boldsymbol{R}_j$ is given as:

$$\boldsymbol{R}_j = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}. \tag{14}$$

Under the null hypothesis $H_0: \boldsymbol{R}_j \cdot \boldsymbol{\beta}_j^* = \mathbf{0}$, W_j is assumed to be asymptotically χ^2 distributed with $2^{K_j+1} - 2 = 6$ degrees of freedom. It should be noted that the Wald test for comparing the reduced and saturated models only requires estimation of the saturated model. That is, finding $\widehat{\boldsymbol{\beta}}_{Rj}$, $\widehat{\boldsymbol{\beta}}_{Fj}$, variance-covariance matrices of $\widehat{\boldsymbol{\beta}}_{Rj}$, $\widehat{\boldsymbol{\beta}}_{Fj}$, and $\boldsymbol{R}_j$ is sufficient to implement the Wald test. The implementation of the Wald test for DIF analysis rests on an important property of the chosen CDM that its item parameters are absolutely invariant. When the model reasonably fits the data, one can expect the chosen CDM to yield relatively invariant item parameter estimates.

2. METHOD

2.1. Real Data Application: Proportional Reasoning Data

The Proportional Reasoning (PR) data consist of responses of 301 students from the reference (R) group and 506 students from the focal (F) group. The Q-matrix for the PR test is given in Table 1. In estimating both the DINA model and G-DINA model parameters, the MMLE algorithm written in Ox (Doornik, 2002) was implemented with a convergence criterion of 0.001. DIF analyses were conducted using the Wald test in conjunction with the DINA model, with the G-DINA model where the underlying restricted model was not specified, and with the G-DINA model where the underlying DINA model was assumed.

Table 1. *Q-matrix for the PR Data*

Item	α_1	α_2	α_3	α_4	α_5	α_6	Item	α_1	α_2	α_3	α_4	α_5	α_6
1	1	0	1	0	0	1	17	1	1	0	0	0	0
2	0	1	1	0	0	0	18	1	0	1	0	1	1
3	1	0	1	1	1	0	19	1	0	1	1	1	0
4	1	1	1	0	0	0	20	0	0	1	1	1	0
5	1	1	1	0	0	0	21	1	0	1	0	0	0
6	0	0	1	0	0	0	22	1	1	0	0	0	0
7	1	0	0	0	0	1	23	1	0	1	0	0	0
8	0	1	1	0	0	0	24	1	0	1	0	1	1
9	0	0	0	1	0	0	25	1	1	1	0	0	0
10	0	1	0	0	0	0	26	0	1	1	0	0	0
11	1	0	0	0	0	0	27	1	0	0	1	1	0
12	0	0	0	0	1	0	28	1	0	1	0	1	1
13	1	0	0	0	1	0	29	1	0	1	0	1	1
14	1	1	1	0	0	0	30	0	0	0	1	0	0
15	0	0	1	0	0	0	31	1	1	0	0	0	0
16	0	0	1	0	0	0							

2.2. Benchmark Simulation for PR Data Analyses

To fully understand the performance of the Wald test for DIF detection in the three procedures and to generate the empirical distribution of the Wald statistics based on the PR data, a benchmark simulation study was performed that mimicked the PR data. In the benchmark study, Type-I error and power of the Wald test were assessed by generating 500 datasets using the estimated values of the item parameters for the R and F groups combined; and the sample sizes for the R and F groups matched those in the real data. The item parameter estimates under the DINA model were used to generate the datasets, provided in Table 2 for the R, F, and the combined groups. It is known that the theoretical power rate of the Wald tests calculated was inflated; that is why, the empirical distributions of the Wald statistic were obtained.

Table 2. *Item Parameter Estimates of the DINA Model for R, F, and Combined Groups*

Item	$N_R = 301$		$N_F = 506$		$N_T = 807$	
	s (*SE*)	*g* (*SE*)	*s* (*SE*)	*g* (*SE*)	*s* (*SE*)	*g* (*SE*)
1	.024 (.012)	.531 (.049)	.190 (.029)	.451 (.031)	.125 (.018)	.469 (.027)
2	.126 (.025)	.670 (.049)	.220 (.036)	.484 (.028)	.169 (.021)	.514 (.025)
3	.015 (.011)	.844 (.034)	.028 (.015)	.745 (.024)	.026 (.009)	.757 (.021)
4	.282 (.034)	.366 (.051)	.400 (.042)	.240 (.024)	.339 (.026)	.253 (.022)
5	.089 (.022)	.568 (.052)	.117 (.029)	.416 (.027)	.127 (.018)	.451 (.025)
6	.064 (.018)	.708 (.054)	.124 (.025)	.384 (.033)	.086 (.015)	.445 (.029)
7	.009 (.017)	.005 (.088)	.201 (.031)	.040 (.044)	.141 (.020)	.054 (.038)
8	.228 (.032)	.447 (.053)	.365 (.041)	.238 (.024)	.291 (.025)	.269 (.022)
9	.246 (.034)	.390 (.077)	.248 (.031)	.229 (.039)	.257 (.022)	.215 (.038)
10	.045 (.014)	.851 (.045)	.017 (.017)	.643 (.028)	.037 (.011)	.695 (.024)
11	.011 (.007)	.972 (.030)	.036 (.011)	.877 (.030)	.024 (.007)	.892 (.025)
12	.518 (.035)	.001 (.107)	.600 (.035)	.222 (.043)	.584 (.024)	.228 (.037)
13	.601 (.034)	.473 (.073)	.504 (.038)	.275 (.031)	.554 (.025)	.311 (.028)
14	.730 (.033)	.345 (.050)	.656 (.040)	.324 (.026)	.709 (.024)	.341 (.024)
15	.053 (.017)	.656 (.057)	.126 (.026)	.454 (.033)	.084 (.015)	.489 (.029)
16	.075 (.021)	.435 (.062)	.229 (.031)	.397 (.033)	.164 (.019)	.412 (.029)
17	.122 (.024)	.507 (.064)	.218 (.037)	.278 (.025)	.161 (.020)	.313 (.024)
18	.014 (.010)	.908 (.027)	.025 (.013)	.594 (.028)	.016 (.007)	.662 (.023)
19	.144 (.030)	.576 (.049)	.271 (.036)	.231 (.025)	.204 (.022)	.278 (.023)
20	.178 (.034)	.286 (.047)	.459 (.040)	.182 (.023)	.346 (.025)	.195 (.021)
21	.253 (.032)	.363 (.056)	.433 (.036)	.158 (.024)	.333 (.024)	.184 (.022)
22	.168 (.028)	.220 (.062)	.481 (.042)	.081 (.015)	.284 (.025)	.086 (.015)
23	.269 (.032)	.380 (.057)	.546 (.036)	.220 (.027)	.405 (.024)	.245 (.024)
24	.326 (.037)	.490 (.047)	.509 (.039)	.277 (.026)	.401 (.026)	.307 (.023)
25	.142 (.027)	.447 (.053)	.320 (.040)	.353 (.026)	.241 (.023)	.372 (.024)
26	.271 (.033)	.305 (.050)	.507 (.042)	.227 (.024)	.378 (.026)	.230 (.021)
27	.106 (.026)	.567 (.056)	.148 (.030)	.276 (.027)	.130 (.019)	.305 (.026)
28	.114 (.025)	.567 (.046)	.142 (.029)	.298 (.027)	.136 (.019)	.353 (.024)
29	.012 (.009)	.800 (.036)	.108 (.027)	.483 (.029)	.040 (.011)	.533 (.025)
30	.427 (.038)	.179 (.068)	.510 (.033)	.104 (.029)	.494 (.023)	.101 (.029)
31	.201 (.029)	.566 (.062)	.274 (.039)	.263 (.025)	.224 (.023)	.306 (.024)

3. RESULT

Results are reported in this section of the paper. In the first part, preliminary results based on PR data were discussed. In the next part, results of a benchmark simulation study to mimick the PR data were presented.

3.1. Preliminary Results: PR Data Analyses

The first Wald test was conducted with the item parameters calibrated along with the restriction matrix formulated in the DINA model. The second Wald test was conducted with the item parameters calibrated along with the restriction matrix formulated in the G-DINA model where no underlying constrained model was specified. The last Wald test was also conducted with the item parameters calibrated by the G-DINA model, but the restriction matrix was formulated in the G-DINA model framework where underlying DINA model was specified.

Table 3. *Preliminary DIF Results for PR Data*

Item	DINA			G-DINA (No Model Assumed)			G-DINA (DINA Model Assumed)		
	Wald Statistic	*p*-value	DIF	Wald Statistic	*p*-value	DIF	Wald Statistic	*p*-value	DIF
1	30.6	0.000	√	–	–	–	–	–	–
2	16.7	0.000	√	22.5	0.000	√	51.3	0.000	√
3	6.6	0.038	–	10.2	0.856	–	604.2	0.000	√
4	10.7	0.005	–	21.2	0.007	–	40.8	0.000	√
5	7.7	0.022	–	26	0.001	–	169.6	0.000	√
6	32.9	0.000	√	33.8	0.000	√	33.8	0.000	√
7	32.1	0.000	√	–	–	–	–	–	–
8	21.8	0.000	√	20.6	0.000	–	25	0.000	–
9	3.9	0.143	–	11.9	0.003	–	11.9	0.003	–
10	15.9	0.000	–	36.4	0.000	√	36.4	0.000	√
11	9.9	0.007	–	7.7	0.022	–	7.7	0.022	–
12	4.7	0.095	–	20.8	0.000	√	20.8	0.000	√
13	8.4	0.015	–	9.9	0.042	–	42.6	0.000	√
14	2.1	0.349	–	28.6	0.000	–	44.3	0.000	√
15	17.3	0.000	√	16	0.000	–	16	0.000	–
16	18.7	0.000	√	29.7	0.000	√	29.7	0.000	√
17	17.8	0.000	√	19.9	0.001	–	73.4	0.000	√
18	67.6	0.000	√	–	–	–	–	–	–
19	52.4	0.000	√	51	0.000		314.5	0.000	√
20	36.4	0.000	√	24	0.002	–	90	0.000	√
21	28.2	0.000	√	21.1	0.000	√	44.4	0.000	√
22	46.2	0.000	√	44.6	0.000	√	62.8	0.000	√
23	43.3	0.000	√	32.1	0.000	√	41.3	0.000	√
24	29.8	0.000	√	–	–	–	–	–	–
25	17.2	0.000	√	27.5	0.001	–	81.1	0.000	√
26	22.8	0.000	√	12.7	0.013	–	18.9	0.004	–
27	25.2	0.000	√	64.5	0.000	√	155.5	0.000	√
28	26.5	0.000	√	–	–	–	–	–	–
29	61.8	0.000	√	–	–	–	–	–	–
30	5.1	0.077	–	2.1	0.343	–	2.1	0.343	–
31	24.3	0.000	√	40.7	0.000	√	49.5	0.000	√

Notes:
1. $\alpha = 0{:}01{=}31$ was used as the critical value because the theoretical χ^2 distribution can lead to inflated Type-I error.
2. For some of the items, the inverse of the weighted variance-covariance matrix cannot be computed.

Results given in Table 3 showed that the Wald test in the G-DINA model where no underlying constrained model was assumed detected the lowest number of DIF items (n = 11), while the Wald test in the DINA model detected the highest number of DIF items (n = 21). The Wald test in the G-DINA model with the DINA model in the restriction matrix detected 19 DIF items. The agreement among the three Wald tests calculated based on the kappa coefficient was 0.18.

3.2. Benchmark Simulation Study

For the Wald test to adhere well to the nominal significance level ($\alpha = 0.05$), the observed Type-I error should be within the range of (0.04, 0.06) based on the exact binomial distribution where the standard error of p was computed as $[p(1-p)/n]^{1/2}$. Additionally, the critical values of the empirical distributions of the Wald statistics were used to calculate the empirical power of the Wald tests in the benchmark power study and to determine the significance of DIF detection in this dataset. A cutoff of 0.80 indicates excellent power and moderate power between 0.70 and 0.80 (Cohen, 1992).

Table 4 summarizes the results of the benchmark simulation. The Wald test to detect DIF in the DINA model adhered well to the nominal significance level for six items (3, 5, 11, 14, 18, and 29). The observed Type-I error were slightly inflated (within the range of [0.06, 0.10]) for eight items (1, 2, 4, 19, 23, 26, 28, and 31). For the most of the other items, the observed Type-I error were largely inflated. For the most of the items, the Wald test had moderate to excellent power. However, for items 1, 7, 9, 12, 14, 16, 20, 25, and 30, empirical power was inadequate. The observed Type-I error were largely inflated to detect DIF in the G-DINA model. For some of the items, the inverse of the weighted variance-covariance matrix cannot be computed therefore the Wald statistic cannot be aquired, noted as “N/A” in the table. For 10 items (2, 6, 8, 11, 19, 21, 22, 23, 27, and 31), the Wald test had moderate to excellent power when it was used to detect DIF in the G-DINA model. While for the Wald test to detect DIF in the G-DINA model with the DINA model assumed as the underlying restricted model, it had moderate to excellent power only for four items (6, 11, 22, and 31). Because of the highly inflated Type-I error and low power in the G-DINA model, the Wald test in the DINA model was selected to detect DIF in the PR data.

Table 5 presents empirical DIF analysis results on the PR data. Critical values of the empirical distributions were used to determine if an item has DIF. As can be seen in Table 5, most of the items showed DIF except for items 9, 12, 13, 14, and 30 in the DINA model. Most of the DIF items in the PR data were also identified as displaying moderate to excellent power, except for items 1, 7, 16, 20, and 25. Hence, one can be sure that these items are DIF items. Among the five non-DIF items in the PR data, only one item 13 displayed excellent power, therefore this item is a non-DIF item. For those nine items displaying poor power, one has to be cautious in interpreting DIF in these items. It is possible that some of these items are DIF items but are not identified as such because the Wald test for DIF detection in the DINA model is not sensitive enough given the characteristics of the data. One of the reasons could be the small sample size. The other reasons including the items with low discriminating power and small DIF sizes also contribute to the low power.

Table 4. *Benchmark Simulation Study Results*

Item	DINA		G-DINA (No Model Assumed)		G-DINA (DINA Model Assumed)	
	Type-I Error	Empirical Power	Type-I Error	Empirical Power	Type-I Error	Empirical Power
1	0.07	0.54	N/A	N/A	N/A	N/A
2	0.10	0.97	0.40	0.71	0.54	0.61
3	0.02	0.90	0.73	0.06	0.96	0.12
4	0.09	0.84	0.64	0.28	0.87	0.28
5	0.05	0.93	0.72	0.21	0.90	0.15
6	0.18	0.98	0.29	0.90	0.29	0.90
7	0.27	0.04	N/A	N/A	N/A	N/A
8	0.11	0.98	0.45	0.71	0.61	0.55
9	0.53	0.15	0.61	0.10	0.61	0.10
10	0.15	0.74	0.25	0.67	0.25	0.67
11	0.06	0.89	0.11	0.79	0.11	0.79
12	0.30	0.16	0.47	0.16	0.47	0.16
13	0.17	0.85	0.57	0.35	0.75	0.21
14	0.06	0.14	0.72	0.08	0.88	0.11
15	0.17	0.83	0.29	0.60	0.29	0.60
16	0.13	0.68	0.22	0.36	0.22	0.36
17	0.11	0.98	0.53	0.64	0.71	0.47
18	0.02	1.00	N/A	N/A	N/A	N/A
19	0.10	1.00	0.81	0.94	0.99	0.12
20	0.21	0.65	0.74	0.32	0.94	0.21
21	0.14	0.99	0.43	0.75	0.60	0.60
22	0.16	0.95	0.41	0.82	0.58	0.79
23	0.10	0.99	0.38	0.79	0.54	0.57
24	0.11	1.00	N/A	N/A	N/A	N/A
25	0.11	0.67	0.74	0.14	0.91	0.12
26	0.10	0.75	0.45	0.29	0.58	0.20
27	0.15	1.00	0.85	0.74	0.97	0.66
28	0.07	1.00	N/A	N/A	N/A	N/A
29	0.04	1.00	N/A	N/A	N/A	N/A
30	0.47	0.15	0.55	0.15	0.55	0.15
31	0.10	1.00	0.52	0.89	0.68	0.76

Table 5. *Empirical DIF Results for PR Data*

Item	DINA			G-DINA (No Model Assumed)			G-DINA (DINA Model Assumed)		
	Wald Statistic	DIF	Power	Wald Statistic	DIF	Power	Wald Statistic	DIF	Power
1	30.60	√	–	–	–	–	–	–	–
2	16.70	√	√	22.50	–	√	51.30	√	–
3	6.60	√	√	10.20	–	–	604.20	√	–
4	10.70	√	√	21.20	–	–	40.80	–	–
5	7.70	√	√	26.00	–	–	169.60	√	–
6	32.90	√	√	33.80	√	√	33.80	√	√
7	32.10	√	–	–	–	–	–	–	–
8	21.80	√	√	20.60	–	√	25.00	–	–
9	3.90	–	–	11.90	–	–	11.90	–	–
10	15.90	√	√	36.40	√	–	36.40	√	–
11	9.90	√	√	7.70	–	√	7.70	–	√
12	4.70	–	–	20.80	–	–	20.80	–	–
13	8.40	–	√	9.90	–	–	42.60	–	–
14	2.10	–	–	28.60	–	–	44.30	–	–
15	17.30	√	√	16.00	√	–	16.00	√	–
16	18.70	√	–	29.70	√	–	29.70	√	–
17	17.80	√	√	19.90	–	–	73.40	√	–
18	67.60	√	√	–	–	–	–	–	–
19	52.40	√	√	51.00	–	√	314.50	√	–
20	36.40	√	–	24.00	–	–	90.00	√	–
21	28.20	√	√	21.10	–	√	44.40	√	–
22	46.20	√	√	44.60	√	√	62.80	√	√
23	43.30	√	√	32.10	√	√	41.30	√	–
24	29.80	√	√	–	–	–	–	–	–
25	17.20	√	–	27.50	–	–	81.10	–	–
26	22.80	√	√	12.70	–	–	18.90	–	–
27	25.20	√	√	64.50	–	√	155.50	√	–
28	26.50	√	√	–	–	–	–	–	–
29	61.80	√	√	–	–	–	–	–	–
30	5.10	–	–	2.10	–	–	2.10	–	–
31	24.30	√	√	40.70	√	√	49.50	–	√

Notes:

1. Power with √ indicates moderate to excellent power, above 0.70.
2. For some of the items, the inverse of the weighted variance-covariance matrix cannot be computed.

There were six attributes in the model. Table 6 lists the estimates of the attribute prevalence for the R and F groups. Among the six listed attributes, Attribute 1 was the easiest one to master for the R group and Attribute 6 was the easiest one for the F group. Attribute 3 was the most difficult one to master for the R group and Attribute 2 was the most difficult one for the F group. Overall, the R group has a higher prevalence of mastering each attribute.

Table 6. *Attribute Prevalence Estimates for the Comparison Groups*

Item	Posterior Probability	
	R	F
1	0.889	0.710
2	0.765	0.368
3	0.725	0.476
4	0.744	0.571
5	0.841	0.596
6	0.802	0.755

4. DISCUSSION and CONCLUSION

Designing assessments in CDMs for diagnostic purposes depends on assurance that the methodological advancement is needed for their analysis and commonly use. The invariance of item parameters for various groups of interest should be checked to assure the appropriate use of CDMs. In this sense, DIF analysis is critical for test validation to investigate whether the groups identified ahead of time influence test inference. This study presents the Wald test to detect DIF in different CDM contexts, including the Wald test in the DINA model, in the G-DINA model where the underlying restricted model was not specified, and in the G-DINA model where the underlying DINA model was assumed. For these purposes, low- versus high-performing school districts based on the Proportional Reasoning test were examined for DIF analyses.

From the preliminary DIF detection results, 11 items were identified as DIF items when the Wald test was used in the G-DINA model; 21 items were identified as DIF items in the DINA model; and 19 items were identified as DIF items when the Wald test was used with the saturated G-DINA model but with the DINA model in the restriction matrix. The kappa coefficient of 0.18 indicated a low agreement among the three Wald tests in determining which items were flagged as DIF items.

In addition to the preliminary DIF analyses, a simulation study was implemented to serve as the benchmark to assess the Type-I error and power of the three Wald tests. The Wald test in the DINA model showed a better performance of detecting DIF than the other two tests in terms of the lower Type-I error and more adequate power overall. Based on the empirical DIF results, the Wald test in the DINA model had moderate to excellent power on 22 items. However, the Wald test in the G-DINA model had moderate to excellent power on 10 items; and the Wald test in the G-DINA model where the DINA model was assumed in the restriction matrix had acceptable power only on four items. Because the proposed Wald tests are based on item parameter estimation, the poor performance of the Wald test in the G-DINA model may relate to the small sample size of the real data in application.

Adding to previous studies of using the Wald test to detect DIF in the DINA model, this study explored different ways of constructing the Wald tests in various CDM context and compared the performance of the Wald tests to detect DIF in each of the three scenarios described above. It also discussed how to implement a benchmark simulation study to assess the Type-I error and power of the Wald test applied to real data. Although the proposed Wald tests in the G-

DINA model framework is not as good as the one in the DINA model given the small sample size of the real data, it provides a different way of constructing the test for DIF detection in a more general theoretical framework and can be used to different data application in the future.

Acknowledgements

An early draft of this paper was presented at the annual meeting of National Council on Measurement in Education, San Antonio, TX, USA. (2017, April).

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID

Likun Hou https://orcid.org/0000-0002-1381-8907
Ragip Terzi https://orcid.org/0000-0003-3976-5054
Jimmy de la Torre https://orcid.org/0000-0002-0893-3863

5. REFERENCES

Camilli, G. (2006). Test fairness. *Educational Measurement, 4*, 221-256.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179-199.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333-353.

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.

Doornik, J. A. (2002). *An object-oriented matrix programming using Ox (Version 3.1) [Computer software]*. London, UK: Timberlake Consultants Press.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301-321.

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

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*, 98-125.

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.

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

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement, 40*, 200-217.

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

Morrison, D. F. (1967). *Multivariate statistical methods.* New York, NY: McGraw-Hill.

Shealy, R., & Stout, W. (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*, 159-194.

Tatsuoka, K. K. (1984). *Analysis of errors in fraction addition and subtraction problems*. Computer-based Education Research Laboratory, University of Illinois.

Terzi, R. (2017). *New Q-matrix validation procedures* (Doctoral dissertation). Rutgers, The State University of New Jersey, New Brunswick, NJ.

Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal, 26*, 237-255.

Zhang, W. (2006). *Detecting differential item functioning using the DINA model* (Doctoral dissertation). University of North Carolina at Greensboro, Greensboro, NC.

Zumbo, B. D. (2007). Three generations of dif analyses: Considering where it has been where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233.