JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
Journal of Educational Measurement
Spring 2014, Vol. 51, No. 1, pp. 98–125

Differential Item Functioning Assessment in Cognitive Diagnostic Modeling: Application of the Wald Test to Investigate DIF in the DINA Model

Likun Hou
American Institute of CPAs
Jimmy de la Torre
Rutgers, The State University of New Jersey
Ratna Nandakumar
University of Delaware

Analyzing examinees' responses using cognitive diagnostic models (CDMs) has the advantage of providing diagnostic information. To ensure the validity of the results from these models, differential item functioning (DIF) in CDMs needs to be investigated. In this article, the Wald test is proposed to examine DIF in the context of CDMs. This study explored the effectiveness of the Wald test in detecting both uniform and nonuniform DIF in the DINA model through a simulation study. Results of this study suggest that for relatively discriminating items, the Wald test had Type I error rates close to the nominal level. Moreover, its viability was underscored by the medium to high power rates for most investigated DIF types when DIF size was large. Furthermore, the performance of the Wald test in detecting uniform DIF was compared to that of the traditional Mantel-Haenszel (MH) and SIBTEST procedures. The results of the comparison study showed that the Wald test was comparable to or outperformed the MH and SIBTEST procedures. Finally, the strengths and limitations of the proposed method and suggestions for future studies are discussed.

Background

Over the past decade, cognitive diagnostic models (CDM) have been actively studied as a psychometric research topic (Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010). These types of models can reveal finer grained information about examinees' mastery of interrelated but separable latent skills (i.e., attributes). These models meet the need for providing more diagnostic information and are more relevant to classroom instruction and learning. This need has been articulated by various stakeholders across the educational sectors in the nation (Huff & Goodman, 2007). Although classifying examinees into multidimensional skill groups is not a novel idea (i.e., scale subscores in the classical test theory or multidimensional item response theory model serve similar purposes), such classification can have poor reliability for most practical test lengths. In contrast, classifications based on CDMs have better reliability than multidimensional IRT models for tests of the same length (Templin & Bradshaw, 2013). The ability of CDMs to provide finer grained diagnostic information can facilitate a better understanding of the nature of an examinee's knowledge. As such, CDMs have the potential to improve the practice of large-scale testing.

At present, study of CDMs is mostly limited to research settings because many psychometric questions about the framework have yet to be addressed. One such question pertains to the important topic of exploring differential item functioning (DIF) in CDMs. In the context of CDMs, DIF is an effect where the probabilities of correctly answering an item are different for examinees with the same attribute mastery profile but who are from different observed groups. In other words, the item responses are not independent when conditioned on the attribute profiles alone. The presence of DIF items in CDMs may result in invalid item parameter estimates for either studied group and distort latent attribute profile estimates. Hence, DIF analysis is necessary to establish parameter or construct invariance or lack thereof (Zumbo, 2007). In the CDM, item parameters represent the interaction between the attributes and the items or the problems. Thus, DIF analysis can shed light on whether the attribute-problem interactions are invariant across groups. The lack of invariance can provide invaluable information about the extent to which group membership can affect how specific problems are perceived and solved. Because group membership can act as a proxy for variables that cause DIF, establishing invariance across groups is a necessary step prior to the application of CDMs in comparative research.

To date, only a few studies have been reported on DIF detection in CDMs (i.e., Li, 2008; Milewski & Baron, 2002; Zhang, 2006). Milewski and Baron examined group differences in skill mastery profiles controlling for overall ability in the CDM framework. The intent of the Milewski and Baron (2002) study was not to explore whether an item was biased in measuring a skill, but rather to investigate skill strengths and weaknesses of certain groups such as students in a particular school compared to a random sample of the population after controlling for the total test score. Zhang (2006) examined DIF in CDMs by applying the Mantel-Haenszel (MH; Holland & Thayer, 1988) and SIBTEST methods (Shealy & Stout, 1993a, 1993b). to the deterministic input, noisy "and" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model. The research investigated the efficiency of the MH and SIBTEST methods with two matching criteria (total test scores and attribute profile scores) for DIF detection in the context of CDMs and found that MH and SIBTEST with attribute profile score as the matching criterion have equivalent or better performance compared to MH and SIBTEST procedures where examinees are matched on total scores. However, the two procedures were only capable of detecting uniform DIF with either matching criterion. Moreover, the item parameters and attribute mastery profiles were estimated on the entire test including potential DIF items for the combined sample of the reference and focal groups. Consequently, these estimates were contaminated due to the presence of DIF items. Li (2008) extended Milewski and Baron's study of differential skill functioning and employed a modified higher order DINA model (de la Torre & Douglas, 2004) to simultaneously detect DIF and differential attribute functioning (DAF). In the present study, DIF is defined as a differential probability of success on the item among the groups of interest conditioned on the attribute mastery profile, whereas DAF is defined as the differential probability of mastering an attribute among the groups of interest matched on the general ability The higher order DINA model provided a natural framework of the hierarchical relationship among items, attributes, and general ability to investigate DIF and DAF simultaneously. However, Li's study has some serious limitations. First, only

uniform DIF was studied. Second, the Type I error rates in some simulation conditions were either overestimated or underestimated. And third, the fit of the model to the data was a cause of concern. These studies indicate that prior work on DIF detection in the context of CDMs falls short on many levels, the most important being the contamination of the item parameter and attribute profile estimates by potential DIF items in the test; this leads to the need for a purification step as part of the DIF detection process. Given these limitations, a more effective method for detecting DIF in CDMs remains a worthwhile topic to explore.

The primary objective of this study is to propose a new model-based DIF detection procedure in the DINA model by adapting the Wald test (Morrison, 1967, p.129) for multivariate hypothesis testing for DIF detection. The same definition of DIF used by Li (2008) is used here. The effectiveness of the Wald test to detect both uniform and nonuniform DIF will be investigated utilizing the DINA model. By formulating the hypothesis test of the joint difference between the item parameters of the reference and focal groups, DIF items can be detected via the Wald test through separate item calibration for the reference and focal groups. This feature is one of the major strengths of the proposed procedure—by performing separate calibrations for the reference and focal groups, the procedure avoids contamination due to DIF items, and thus obviates the need for a test purification process. The procedure also has the potential advantage of detecting both uniform and nonuniform DIF efficiently. The viability of the proposed method was assessed through a simulation study by investigating its Type I error and power. A variety of factors (sample size, reference group parameter values, number of attributes required by the item, DIF size, and DIF type) were examined to establish the generalizability of the simulation results. The proposed method also was compared with the traditional MH and SIBTEST procedures in detecting uniform DIF.

Theoretical Framework

The DINA Model

The DINA model is one of the most parsimonious CDMs for dichotomously scored test items. This model is appropriate when all the required attributes for an item have to be mastered to answer an item correctly, and lacking one or more required attributes for an item is the same as lacking all of them. In this model, each item has two parameters regardless of the number of attributes required to answer it correctly.

Let X_{ij} be the response of examinee i ($i = 1, \ldots, I$) to item j ($j = 1, \ldots, J$), and let $\boldsymbol{\alpha}_i = \{\alpha_{ik}\}$ be the examinee's binary attribute vector, $k = 1, \ldots, K$, where $\alpha_{ik} = 1$ indicates mastery of attribute k by examinee i and 0 indicates nonmastery of the attribute by the examinee. The Q-matrix (Tatsuoka, 1983) is constructed as a $J \times K$ item-attribute matrix of zeros and ones with element $q_{jk} = 1$ denoting that attribute k is required to correctly answer item j, and $q_{jk} = 0$, and the attribute is not required for the item. In the DINA model, the probability of correctly answering item j for examinee i with the attribute profile $\boldsymbol{\alpha}_i$ is represented by:

$$P_j(\boldsymbol{\alpha}_i) = P(X_{ij} = 1|\boldsymbol{\alpha}_i) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}, \tag{1}$$

where

$$\eta_{ij} = \prod_{k}^{K} \alpha_{ik}^{q_{jk}}. \quad (2)$$

In both Equations 1 and 2, η_{ij} assumes the value 1 or 0, indicating whether the examinee i with the attribute profile $\boldsymbol{\alpha}_i$ has mastered all of the required attributes ($\eta_{ij} = 1$) or lacks at least one of the required skills ($\eta_{ij} = 0$) for item j. The slip and guessing parameters of item j are defined as $s_j = P(X_{ij} = 0|\eta_{ij} = 1)$ and $g_j = P(X_{ij} = 1|\eta_{ij} = 0)$, respectively. Once η_{ij} has been determined, the probability that examinee i answers item j correctly is $1 - s_j$ if $\eta_{ij} = 1$, and g_j if $\eta_{ij} = 0$. To obtain the parameter estimates of the DINA model, the marginalized maximum likelihood estimation (MMLE) method can be implemented using the EM algorithm (Dempster, Laird, & Rubin, 1977). The details of the EM algorithm for estimating the DINA model parameters and their corresponding standard errors can be found in de la Torre (2009).

DIF in Cognitive Diagnostic Modeling

Traditional DIF detection procedures in the context of IRT models investigate whether examinees that have the same latent ability level or total test score but are from different groups have a different probability of endorsing an item. It has been shown that the presence of DIF items on the test can lead to invalid item parameter estimates and distort latent trait estimates that could contribute to test unfairness for examinees from one of the groups (Lord, 1980). DIF can be equally important in the context of CDMs. For CDMs, DIF needs to be redefined because the examinees are provided with the mastery profile of latent attributes instead of a rank on the latent ability level. An item exhibits DIF in the context of CDMs if the probabilities of success on the item are different for examinees that have the same attribute mastery profile but are from different groups.

DIF in CDMs can be represented as $\Delta_{j\boldsymbol{\alpha}_l} = P(X_j = 1\,|\boldsymbol{\alpha}_l)_{\mathrm{F}} - P(X_j = 1\,|\boldsymbol{\alpha}_l)_{\mathrm{R}}$, where $\Delta_{j\boldsymbol{\alpha}_l}$ denotes DIF in item j for examinees with the attribute mastery profile $\boldsymbol{\alpha}_l$; $P(X_j = 1|\boldsymbol{\alpha}_l)_{\mathrm{F}}$ is the success probability on item j for examinees with the attribute mastery profile $\boldsymbol{\alpha}_l$ in the focal group; and $P(X_j = 1|\boldsymbol{\alpha}_l)_{\mathrm{R}}$ is the success probability on item j for examinees with the attribute mastery profile $\boldsymbol{\alpha}_l$ in the reference group. When $\Delta_{j\boldsymbol{\alpha}_l} > 0$, DIF is against the reference group and the item favors the focal group. On the other hand, when $\Delta_{j\boldsymbol{\alpha}_l} < 0$, DIF is against the focal group and the item favors the reference group. No DIF is detected in the item if $\Delta_{j\boldsymbol{\alpha}_l} = 0$ for all attribute mastery profiles.

Similar to DIF in the context of IRT, uniform and nonuniform DIF can be detected in CDMs. Uniform DIF exists if the probabilities of correctly answering the item are consistently lower or higher for one group regardless of the latent attribute profile, that is, $\Delta_{j\boldsymbol{\alpha}_l}$ is either positive or negative for all latent attribute profiles. Nonuniform DIF exists if the probabilities of correctly answering the item are lower for one group on some latent attribute profiles but higher for the same group on some other latent attribute profiles. That is, $\Delta_{j\boldsymbol{\alpha}_l}$ changes its sign depending upon the attribute profiles.

In the DINA model there are two parameters (the slip and guessing parameters) for each item associated with two attribute mastery categories. Item j exhibits DIF

if:

$$\Delta_{j,\eta_j=1} = \Delta_{sj} = P(X_{ij} = 1|\eta_{ij} = 1)_{\mathrm{F}} - P(X_{ij} = 1|\eta_{ij} = 1)_{\mathrm{R}} = s_{\mathrm{R}j} - s_{\mathrm{F}j} \neq 0 \tag{3}$$

and/or

$$\Delta_{j,\eta_j=0} = \Delta_{gj} = P(X_{ij} = 1|\eta_{ij} = 0)_{\mathrm{F}} - P(X_{ij} = 1|\eta_{ij} = 0)_{\mathrm{R}} = g_{\mathrm{F}j} - g_{\mathrm{R}j} \neq 0. \tag{4}$$

Thus, whether DIF is present in an item can be detected by examining the differences in the slip and guessing parameters between the focal and reference groups in the DINA model. Uniform DIF occurs in item j when Δ_{sj} and Δ_{gj} have the same signs:

$$\begin{cases} \Delta_{sj} > 0 \quad \text{or} \quad s_{\mathrm{F}j} - s_{\mathrm{R}j} < 0 \\ \Delta_{gj} > 0 \quad \text{or} \quad g_{\mathrm{F}j} - g_{\mathrm{R}j} > 0 \end{cases} \tag{5}$$

or

$$\begin{cases} \Delta_{sj} < 0 \quad \text{or} \quad s_{\mathrm{F}j} - s_{\mathrm{R}j} > 0 \\ \Delta_{gj} < 0 \quad \text{or} \quad g_{\mathrm{F}j} - g_{\mathrm{R}j} < 0. \end{cases} \tag{6}$$

As shown in Equations 5 and 6, uniform DIF is present in item j when the slip parameter is smaller and the guessing parameter is larger for the focal group; this results in higher probabilities of correctly answering the item for examinees in the focal group compared to those in the reference group regardless of their attribute mastery profile, indicating that the item favors the focal group. Another way to produce uniform DIF is when the slip parameter is larger and the guessing parameter is smaller for the focal group; this results in lower probabilities of correctly answering the item for examinees in the focal group compared to those in the reference group regardless of their attribute mastery profile, indicating that the item favors the reference group.

Nonuniform DIF occurs in item j when Δ_{sj} and Δ_{gj} have different signs:

$$\begin{cases} \Delta_{sj} > 0 \quad \text{or} \quad s_{\mathrm{F}j} - s_{\mathrm{R}j} < 0 \\ \Delta_{gj} < 0 \quad \text{or} \quad g_{\mathrm{F}j} - g_{\mathrm{R}j} < 0 \end{cases} \tag{7}$$

or

$$\begin{cases} \Delta_{sj} < 0 \quad \text{or} \quad s_{\mathrm{F}j} - s_{\mathrm{R}j} > 0 \\ \Delta_{gj} > 0 \quad \text{or} \quad g_{\mathrm{F}j} - g_{\mathrm{R}j} > 0. \end{cases} \tag{8}$$

As shown in Equations 7 and 8, nonuniform DIF is present in item j when both the slip and guessing parameters are smaller for the focal group; this results in higher probabilities of success on item j for masters and lower probabilities of success on item j for nonmasters in the focal group compared to those in the reference group, respectively, indicating that the item favors the focal group for the masters yet favors the reference group for the nonmasters. Another way to produce nonuniform DIF is when both the slip and guessing parameters are larger for the focal group; this results

in lower probabilities of success on item j for masters and higher probabilities of success on item j for nonmasters in the focal group compared to those in the reference group, indicating that the item favors the reference group for the masters yet favors the focal group for the nonmasters.

Wald Test for DIF Detection

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

$$H_0 : \begin{cases} \Delta_{sj} = 0 \\ \Delta_{gj} = 0 \end{cases}, \quad \text{or} \quad \begin{cases} s_{\mathrm{F}j} - s_{\mathrm{R}j} = 0 \\ g_{\mathrm{F}j} - g_{\mathrm{R}j} = 0. \end{cases} \tag{9}$$

The alternative hypothesis is: at least one of the item parameters is different between the focal and reference groups. There are two steps to implement the Wald test. In the first step, item parameters are estimated separately for the focal and reference groups (i.e., separate calibrations for the focal and reference groups) using a general formulation of the attribute distribution. The first step translates into applying an unconstrained model to the data, where the item parameters are not set to be equal across the focal and reference groups. The parameter estimates for item j across the two groups are represented as

$$\hat{\boldsymbol{\beta}}_j^* = (\boldsymbol{\beta}_{\mathrm{R}j}, \boldsymbol{\beta}_{\mathrm{F}j}) = (g_{\mathrm{R}j}, s_{\mathrm{R}j}, g_{\mathrm{F}j}, s_{\mathrm{F}j})'. \tag{10}$$

In the second step, the null hypothesis of the equality of item parameters of the focal and reference groups is tested. The second step translates into applying a constrained model to the data, where the item parameters are set to be equal across the focal and reference groups, and examining whether such a constrained model leads to deterioration of the model-data fit (de la Torre, 2011).

The null hypothesis given in Equation 9 can be expressed in terms of the constrained model as follows:

$$H_0 : \mathbf{R}_j \cdot \boldsymbol{\beta}_j^* = \mathbf{0}, \tag{11}$$

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

$$\mathbf{R}_j = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}. \tag{12}$$

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

$$W_j = [\mathbf{R}_j \cdot \boldsymbol{\beta}_j^*]' \{\mathbf{R}_j \cdot \mathrm{Var}(\boldsymbol{\beta}_j^*) \cdot \mathbf{R}_j'\}^{-1} [\mathbf{R}_j \cdot \boldsymbol{\beta}_j^*], \tag{13}$$

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

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_j^*) = \begin{pmatrix} \mathrm{Var}(\widehat{\boldsymbol{\beta}}_{\mathrm{R}j}) & \mathbf{0} \\ \mathbf{0} & \mathrm{Var}(\widehat{\boldsymbol{\beta}}_{\mathrm{F}j}) \end{pmatrix}. \tag{14}$$

Under the null hypothesis that $\mathbf{R}_j \cdot \boldsymbol{\beta}_j^* = \mathbf{0}$, W_j is asymptotically distributed as a chi-square distribution with two degrees of freedom. It should be noted that the Wald test for comparing the constrained model and unconstrained model only requires estimation of the unconstrained model. That is, finding $\widehat{\boldsymbol{\beta}}_{\mathrm{R}j}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{F}j}$, $\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{\mathrm{R}j})$, $\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{\mathrm{F}j})$, and $\mathbf{R}_j$ is sufficient to implement the proposed DIF detection procedure. An important advantage of using separate calibrations for the reference and focal groups is that the procedure obviates the need for an additional step to purify the test to be DIF free prior to item calibration. This additional purification step in itself can be a very involved and contentious process. With separate calibrations, contamination of item parameter and attribute estimates due to items misidentified as non-DIF items becomes a nonissue. As such, the proposed method is relatively efficient and straightforward. The implementation of the Wald test for DIF analysis rests on an important property of the DINA model: its parameters are absolutely invariant. De la Torre and Lee (2010) have shown that unbiased estimates of the DINA model parameters can be obtained regardless of the generating attribute distribution as long as the estimation algorithm uses an attribution that is generally formulated (i.e., unstructured). With real data, a certain degree of model misfit can be expected. However, one can still expect the DINA model to yield relatively invariant item parameter estimates provided that the model reasonably fits the data.

Type I Error and Power Study

The goal of the Type I error and power study was to assess the viability of the proposed Wald test in the DINA model through a simulation study where a variety of factors were examined. The performance of the Wald test in detecting both uniform and nonuniform DIF items was examined.

Design

The performance of traditional DIF detection methods has been shown to be influenced by factors such as sample size, test length, proportion of items in the test exhibiting DIF, the size of DIF, and group impact on ability distribution (Mazor, Clauser, & Hambleton, 1992; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). In addition to these factors, the number of attributes included in the model, the complexity of the Q-matrix, and the values for the slip and guessing parameters also are appropriate to consider in the context of the DINA model. In this simulation study, data were generated using a fixed number of attributes ($K = 5$) and a fixed test length ($J = 30$). The Q-matrix is given in Table 1, which was constructed to be balanced such that success on each item requires either a single attribute, two attributes, or three attributes; and each attribute influences the same number of items as other attributes, resulting in 10 single-attribute items ($K_j = 1$), 10 two-attribute items ($K_j = 2$), and 10 three-attribute items ($K_j = 3$). The joint distributions of attribute profiles were generated with equal probabilities from a multinomial distribution.

Four factors were manipulated to study their influence on DIF detection: sample size for the reference and focal group (N_{R} and N_{F}), reference group item parameter values ($g_{\mathrm{R}j}$ and $s_{\mathrm{R}j}$), DIF size, and DIF type. Sample size consistently is shown to be

Table 1
Q-Matrix for the Simulated Data

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

an important factor in DIF detection procedures. Two sample sizes, 500 and 1,000, were used for each group to represent small and large sample sizes. De la Torre and Lee (2010) reported good recovery of the item parameters for the DINA model with these two sample sizes. In addition, these two sizes were commonly used for other traditional DIF methodologies including the MH and SIBTEST procedures. For example, Zhang (2006) simulated equal sample size of 400 and 800 for both studied groups, respectively. Li (2008) simulated equal sample sizes of 500 and 1,000 for both studied groups. If the Wald test is to be comparable with these methods, it also must show considerable power at comparable sample sizes.

Slip and guessing parameters are both indicators of amount of noise in the data. Smaller values for these parameters would indicate that the item is more capable of distinguishing masters and nonmasters (Templin & Henson, 2006). In the simulation study and the real data example presented in de la Torre and Douglas (2004), the range of most slip and guessing parameters was from .1 to .3. These ranges were incorporated into this study and the slip and guessing parameter values were fixed to be equal for the reference group. Three sets of values were considered: $g_{Rj} = s_{Rj} =$ .1, .2, or .3.

The slip and the guessing parameter values for the focal group were manipulated according to the DIF size (*DS*) and DIF type (*DT*). Two values were considered for DIF size, which is defined as the differences in the guessing parameters or the slip parameters between the focal and reference groups. In this study, two levels of DIF size similar to Zhang's (2006) study were chosen: small ($|\Delta_{sj}|$ or $|\Delta_{gj}| = .05$) and large ($|\Delta_{sj}|$ or $|\Delta_{gj}| = .10$).

For each DIF size, nine possible combinations of Δ_{sj} and Δ_{gj} were considered. For the small DIF size, $\Delta_{sj} = -.05, 0, .05$, and $\Delta_{gj} = -.05, 0, .05$ were completely crossed; for the large DIF size, $\Delta_{sj} = -.10, 0, .10$, and $\Delta_{gj} = -.10, 0, .10$ again were completely crossed. For each DIF size, these combinations resulted in three DIF categories: non-DIF, uniform DIF, and nonuniform DIF. When both Δ_{sj} and Δ_{gj} are equal to zero, it results in the non-DIF category. When Δ_{sj} and Δ_{gj} have the same sign, DIF is uniform. When Δ_{sj} and Δ_{gj} have the opposite sign or one of them is zero, DIF is nonuniform. There are two DIF types in the uniform category and six DIF types in the nonuniform category for each DIF size. The different simulated DIF conditions are summarized in Table 2.

For other DIF detection methods (e.g., the Mantel-Haenszel procedure), the proportion of DIF items in the test has an impact on DIF detection: Too many DIF items can contaminate the conditioning variables (Narayanon & Swaminathan, 1996). However, with our proposed method, item parameters are estimated separately for the reference and focal groups and the difference in the item parameters between the studied groups are tested through the Wald test. Hence, the item parameters and attribute profile estimates are not distorted by the existence of DIF items because estimation is not done on the combined sample. Consequently, the proportion of DIF items does not influence the DIF detection rate of the proposed method. This is further documented in the second study comparing the Wald test and traditional MH and SIBTEST procedures. To simplify the simulation process and separate the conditions for the Type I error study and power study, all 30 items were simulated as having DIF under each DIF condition.

For the Type I error component of the study (i.e., the case of non-DIF), the simulated factors of sample size (2) and item parameter values (3) were completely crossed resulting in six different conditions. For the power component of the study, both uniform and nonuniform DIF categories were evaluated. In the case of uniform DIF, for the small DIF size of .05, the simulated factors of sample size (2), reference item parameter values (3), and DIF type (2) were completely crossed resulting in 12 conditions; for the large DIF size of .10, the simulated factors of sample size (2), reference item parameter values (2), and DIF type (2) were crossed resulting in 8 conditions. Note that for the large DIF size of .10, only the values of .2 and .3 for the

Table 2
Summary of DIF Conditions for the Type I and Power Study

DIF Type	Size	Δ_{sj}	Δ_{gj}
Non-DIF	n/a	0	0
Uniform DIF	Small	+.05	+.05
		−.05	−.05
	Large	+.10	+.10
		−.10	−.10
Nonuniform DIF	Small	+.05	−.05
		−.05	+.05
		+.05	0
		0	+.05
		−.05	0
		0	−.05
	Large	+.10	−.10
		−.10	+.10
		+.10	0
		0	+.10
		−.10	0
		0	−.10

reference item parameters were considered because DIF size has to be less than the reference item parameter values. Similarly, there were 60 different conditions in the case of nonuniform DIF. In total, there were 86 different conditions included in this study.

Examinee responses were generated based on the DINA model using the examinee attribute mastery profiles and the given sets of slip and guessing parameters. Under each condition, one thousand datasets were simulated and the Wald test was applied for all items of each dataset to detect DIF. In each condition, the percentage of replications where H_0 was rejected for each item was observed. The Type I error and power of the Wald test were determined using the significance level of .05. The codes for the data generation, estimation, and DIF analysis were written in Ox (Doornik, 2002).

Results

Type I Error Study

Type I error occurs when an item is detected as having DIF when in fact it does not exhibit DIF toward either group. For each of the 30 items on the test, the Type I error rate is defined as the percentage of times the item was detected as displaying DIF out of the 1,000 replications under each non-DIF condition. The average Type I error rates are reported for one-, two-, and three-attribute items separately.

Table 3 summarizes the Type I error rates of the Wald test as a function of sample size, reference item parameter values, and number of attributes required for success

Table 3
Type I Error Rates for Different Non-DIF Conditions (α = .05)

Reference Item Parameter Values	Sample Size					
	$N_R = 500, N_F = 500$			$N_R = 1{,}000, N_F = 1{,}000$		
	$K_j = 1$	$K_j = 2$	$K_j = 3$	$K_j = 1$	$K_j = 2$	$K_j = 3$
$g_{Rj} = s_{Rj} = 0.1$	.061	.057	.053	.056	.054	.053
$g_{Rj} = s_{Rj} = 0.2$	.079	.068	.064	.071	.063	.058
$g_{Rj} = s_{Rj} = 0.3$	.182	.145	.113	.135	.106	.083

Note. g_{Rj} = guessing parameter for the reference group; s_{Rj} = slip parameter for the reference group; K_j = number of attributes required for success on the item.

on the item. Because α = .05, the observed Type I error rates would be expected to fall between .04 and .06 95% of the time (based on the exact binomial distribution where the standard error of p was computed as $[p(1 - p)/n]^{1/2}$) if the test were adhering well to the nominal level of .05. It can be seen in Table 3 that the degree of overestimation of the Type I error rates varied. As the sample size increased, the Type I error rates got closer to the nominal level as the estimation error became smaller. As the reference item parameter values decreased from .3 to .1, the Type I error rates got closer to the nominal level. The Type I error rates were more inflated for the items with larger reference item parameter values. In addition, as the number of attributes required for success on the item increased, the Type I error rates were closer to the nominal level. In sum, the Type I error rates were closer to the nominal level as the item slip and guessing parameters became smaller. The Type I error rates were close to the nominal level for the items with small slip and guessing parameter values (.1, and .2), suggesting that the Wald test is an effective and valid procedure to detect DIF in the DINA model when the items are relatively discriminating (i.e., the item guessing and slip parameters are small) even for the small sample size of 500. Generally speaking, when the slip and guessing parameters are large, the item is less capable of distinguishing the masters from nonmasters and the data will be noisier. This in turn negatively affects the ability of the Wald test to detect DIF.

Power Study

In this study, power rates were determined in two ways. The first method was the theoretical value, which was based on the chi-square distribution with 2 degrees of freedom. The second method was based on an empirical distribution for each combination of simulated factors in the non-DIF condition. For each combination of simulated factors (reference item parameter values, sample size, and the number of attributes required for success on the item), an empirical distribution under the null hypothesis was generated and used as the baseline comparison for determining the power. Each distribution included 10,000 sample statistics. The critical values of the empirical distributions of the Wald statistic at α level of .05 are listed in Table 4. The theoretical chi-squared critical value for 2 degrees of freedom at the significance level of .05 is 5.99 for all simulated conditions, whereas the empirical critical values,

Table 4
Critical Values of the Empirical Wald Statistic Distributions (α = .05)

Reference Parameter Values	Sample Size					
	$N_R = 500, N_F = 500$			$N_R = 1{,}000, N_F = 1{,}000$		
	$K_j = 1$	$K_j = 2$	$K_j = 3$	$K_j = 1$	$K_j = 2$	$K_j = 3$
$g_{Rj} = s_{Rj} = 0.1$	6.38	6.25	6.06	6.22	6.16	6.10
$g_{Rj} = s_{Rj} = 0.2$	7.08	6.67	6.60	6.86	6.42	6.33
$g_{Rj} = s_{Rj} = 0.3$	11.08	9.37	8.43	8.93	8.02	7.35

Note. g_{Rj} = guessing parameter for the reference group; s_{Rj} = slip parameter for the reference group; K_j = number of attributes required for success on the item.

in Table 4, range from 6.06 to 11.08 and differ according to the combination of simulated factors.

Power rate is defined as the percentage of DIF items correctly detected out of 1,000 replications. The average power rate was reported for one-, two-, and three-attribute items respectively for each DIF condition. A cutoff of .80 was used to indicate excellent power and power rates between .70 and .80 were evaluated as moderate (Cohen, 1992).

Detecting uniform DIF. Tables 5 and 6 list the power rates of the Wald test to detect uniform DIF calculated using the theoretical χ^2 ($df = 2$) distribution and using the empirical distributions of the Wald statistic respectively. It should be noted that the interpretation of theoretical power rates is conditional on the Type I error rates for a given significance level because the power rates can artificially increase if the Type I error rates are inflated. Compared to the theoretical power rates, the empirical power rates were slightly lower. For example, as listed in Tables 5 and 6, when reference item slip and guessing parameters = .1, DIF size of .05 and sample size of 500 for either studied group, in the case of DIF against the reference group, the average theoretical power rates versus the average empirical power rates were .614 versus .582 for single-attribute items, .598 versus .574 for two-attribute items, and .572 versus .566 for three-attribute items. The same trend holds for each combination of the simulated factors. To be conservative, comparison of power rates across the simulated conditions is conducted using only the empirical power rates.

As can be seen in Table 6, power rates increased as the sample size increased irrespective of other factors. For example, when the reference item slip and guessing parameters were equal to .2 and DIF size was .05, comparing the sample size of 1,000 to 500 for either studied group, the power rates were 59.0% higher for the sample size of 1,000 in single-attribute items, 66.9% higher in two-attribute items, and 66.6% higher in three-attribute items on average.

The reference item parameter values also influenced the power rates. As the reference item parameter values increased, the power rates decreased. Figures 1 and 2 plot the power rates averaged over the number of attributes required by the item by the reference item parameter value for each uniform DIF type against the reference or focal groups, respectively, when the DIF size was .05 and sample size was

Table 5
Power Rates for Detecting Uniform DIF based on the χ^2 ($df = 2$) Distribution

			Sample Size							
			$N_R = 500, N_F = 500$				$N_R = 1{,}000, N_F = 1{,}000$			
g_{Rj}/s_{Rj}	DIF Size	DIF Type	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
.1	.05	$\Delta_{sj} = +0.05, \Delta_{gj} = +0.05$	.614	.598	.572	.595	.901	.887	.891	.893
		$\Delta_{sj} = -0.05, \Delta_{gj} = -0.05$	.621	.713	.757	.697	.906	.955	.972	.944
.2	.05	$\Delta_{sj} = +0.05, \Delta_{gj} = +0.05$	.390	.382	.394	.389	.613	.637	.656	.635
		$\Delta_{sj} = -0.05, \Delta_{gj} = -0.05$	.388	.409	.435	.411	.624	.683	.725	.677
	.10	$\Delta_{sj} = +0.10, \Delta_{gj} = +0.10$	.907	.905	.917	.910	.996	.997	.998	.997
		$\Delta_{sj} = -0.10, \Delta_{gj} = -0.10$	.912	.957	.978	.949	.996	1.000	1.000	.999
.3	.05	$\Delta_{sj} = +0.05, \Delta_{gj} = +0.05$	.372	.354	.347	.358	.506	.506	.536	.516
		$\Delta_{sj} = -0.05, \Delta_{gj} = -0.05$	.381	.372	.368	.374	.500	.532	.575	.536
	.10	$\Delta_{sj} = +0.10, \Delta_{gj} = +0.10$	.768	.796	.833	.799	.950	.968	.983	.967
		$\Delta_{sj} = -0.10, \Delta_{gj} = -0.10$	.774	.831	.883	.829	.954	.982	.993	.976

Note. g_{Rj} = guessing parameter for the reference group; s_{Rj} = slip parameter for the reference group; K_j = number of attributes required for success on the item; $\Delta_{gj} = g_{Fj} - g_{Rj}$; $\Delta_{sj} = s_{Rj} - s_{Fj}$.

Table 6
Power Rates for Detecting Uniform DIF based on the Empirical Wald Statistic Distribution

			Sample Size							
			$N_R = 500, N_F = 500$				$N_R = 1{,}000, N_F = 1{,}000$			
g_{Rj}/s_{Rj}	DIF Size	DIF Type	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
.1	.05	$\Delta_{sj} = +0.05, \Delta_{gj} = +0.05$	.582	.574	.566	.574	.893	.879	.886	.886
		$\Delta_{sj} = -0.05, \Delta_{gj} = -0.05$	.592	.693	.751	.679	.900	.953	.970	.941
.2	.05	$\Delta_{sj} = +0.05, \Delta_{gj} = +0.05$	.316	.333	.348	.332	.551	.602	.630	.594
		$\Delta_{sj} = -0.05, \Delta_{gj} = -0.05$	.317	.357	.386	.353	.561	.651	.702	.638
	.10	$\Delta_{sj} = +0.10, \Delta_{gj} = +0.10$	.868	.880	.898	.882	.994	.996	.998	.996
		$\Delta_{sj} = -0.10, \Delta_{gj} = -0.10$	.876	.943	.969	.929	.995	.999	.999	.998
.3	.05	$\Delta_{sj} = +0.05, \Delta_{gj} = +0.05$	.157	.187	.204	.183	.346	.376	.440	.387
		$\Delta_{sj} = -0.05, \Delta_{gj} = -0.05$	.164	.200	.219	.194	.336	.398	.470	.401
	.10	$\Delta_{sj} = +0.10, \Delta_{gj} = +0.10$	.543	.622	.706	.624	.892	.933	.970	.932
		$\Delta_{sj} = -0.10, \Delta_{gj} = -0.10$	.534	.679	.781	.665	.898	.958	.988	.948

Note. g_{Rj} = guessing parameter for the reference group; s_{Rj} = slip parameter for the reference group; K_j = number of attributes required for success on the item; $\Delta_{gj} = g_{Fj} - g_{Rj}$; $\Delta_{sj} = s_{Rj} - s_{Fj}$.

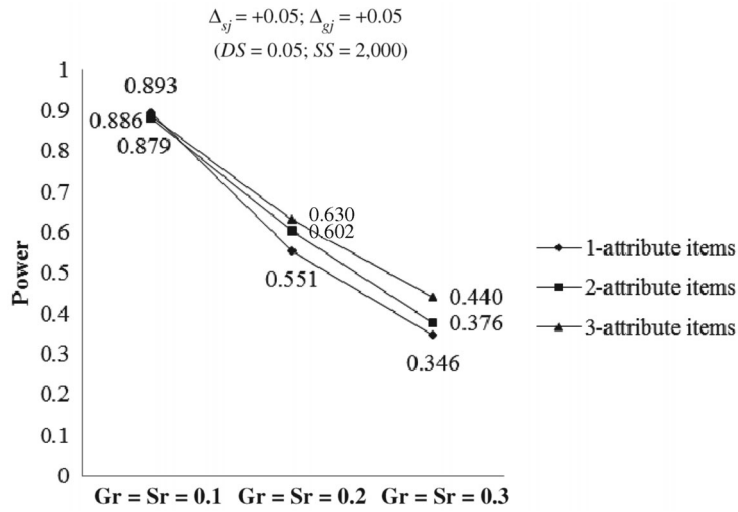

Figure 1. Average empirical power rates by reference item parameter values for uniform DIF against the reference group with DIF size = .05 (N_R = 1,000, N_F = 1,000).

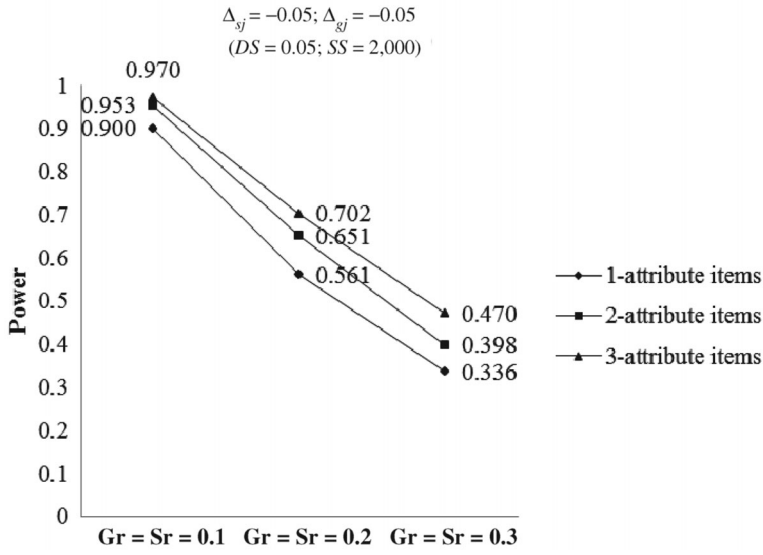

Figure 2. Average empirical power rates by reference item parameter values for uniform DIF against the focal group with DIF size = .05 (N_R = 1,000, N_F = 1,000).

1,000 for either studied group. As can be seen in these two figures, the power rates when the reference item parameters were equal to .3 were less than half of those when the reference item parameters were equal to .1. For example, Figure 1 shows that for single-attribute items exhibiting DIF against the reference group, the power rates were .346 when the reference item parameter values were .3 compared to the power

rates of .893 when the reference item parameters were .1. The trend was consistent across each uniform DIF type, sample size, and the number of attributes required by the item.

DIF size also had a distinctive impact on the power rates. The larger DIF size corresponded to higher power rates across the uniform DIF types and sample sizes. Even when the items were relatively less discriminating ($g_{\mathrm{R}j} = s_{\mathrm{R}j} = .3$), excellent power rates were achieved for the large DIF size of .10 when the sample size was as large as 1,000 for either studied group, and moderate power rates were achieved for the large DIF size of .10 when the sample size was as small as 500 for either group. Overall, when the item was highly discriminating (.1) or DIF size was large (.10), excellent power rates were achieved for detecting uniform DIF for the large sample size of 1,000 for either studied group and moderate to excellent power rates were achieved for the small sample size for either group.

Detecting nonuniform DIF. Tables 7 and 8 list power rates of the Wald test to detect nonuniform DIF calculated using the theoretical χ^2 ($df = 2$) distribution and using the empirical distributions of the Wald statistic respectively. Similar to the performance when detecting uniform DIF, the empirical power rates of the Wald test were slightly lower than the corresponding theoretical power rates when detecting nonuniform DIF. As before, to be conservative, the empirical power rates were used to make comparisons across the simulated conditions for nonuniform DIF detection.

As shown in Table 8, the power rates for detecting nonuniform DIF increased as the sample size increased. Similar to uniform DIF detection, the power rates for detecting nonuniform DIF decreased as the reference item parameter values increased. This is plotted in Figure 3 for a DIF size of .05 and sample size of 1,000 for either studied group. The trend is consistent for the other combinations of DIF size and sample size. In addition, the larger DIF size was associated with the higher power rates across different combinations of the nonuniform DIF types, sample sizes and reference item parameter values. As an example, Figure 4 shows this trend when the reference item parameter values were .2 and the sample size was 500 for either studied group.

In addition, the power rates for detecting nonuniform DIF varied across the DIF types. For the DIF types that involved both masters and nonmasters ($\Delta_{sj} \neq 0$ and $\Delta_{gj} \neq 0$), the power rates were relatively higher compared to those when the DIF types involved only masters or nonmasters ($\Delta_{sj} = 0$ or $\Delta_{gj} = 0$). The power rates for detecting nonuniform DIF that involved only masters were lower than the power rates for the other nonuniform DIF types. This trend was more apparent for the small sample size of 500 for either studied group. Similar to uniform DIF detection, for both small and large sample sizes, the Wald test achieved medium to excellent power when the DIF size was large (.10) or the item was highly discriminating (.1) for the nonuniform DIF type that involved both masters and nonmasters. But for the small DIF size of .05, low discriminating items, and the nonuniform DIF type that involved only masters, the Wald test was not sensitive enough to detect DIF items; this was true for both small and large sample sizes.

Table 7
Power Rates for Detecting Nonuniform DIF based on the χ^2 ($df = 2$) Distribution

			Sample Size							
			$N_R = 500, N_F = 500$				$N_R = 1{,}000, N_F = 1{,}000$			
g_{Rj}/s_{Rj}	DIF Size	DIF Type	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
.1	.05	$\Delta_{sj} = +0.05, \Delta_{gj} = -0.05$	.714	.756	.761	.744	.956	.970	.977	.968
		$\Delta_{sj} = -0.05, \Delta_{gj} = +0.05$	.468	.525	.552	.515	.773	.820	.850	.814
		$\Delta_{sj} = +0.05, \Delta_{gj} = 0$	.415	.216	.101	.244	.717	.421	.233	.457
		$\Delta_{sj} = 0, \Delta_{gj} = +0.05$	.287	.433	.500	.407	.499	.723	.809	.677
		$\Delta_{sj} = -0.05, \Delta_{gj} = 0$	.285	.164	.108	.186	.502	.279	.163	.315
		$\Delta_{sj} = 0, \Delta_{gj} = -0.05$	.418	.627	.723	.589	.714	.914	.957	.861
.2	.05	$\Delta_{sj} = +0.05, \Delta_{gj} = -0.05$	.342	.392	.428	.387	.579	.662	.722	.654
		$\Delta_{sj} = -0.05, \Delta_{gj} = +0.05$	.279	.313	.359	.317	.435	.519	.603	.519
		$\Delta_{sj} = +0.05, \Delta_{gj} = 0$	.222	.128	.094	.148	.367	.193	.129	.230
		$\Delta_{sj} = 0, \Delta_{gj} = +0.05$	.208	.296	.332	.279	.314	.487	.590	.463
		$\Delta_{sj} = -0.05, \Delta_{gj} = 0$	.213	.132	.107	.151	.306	.174	.114	.198
		$\Delta_{sj} = 0, \Delta_{gj} = -0.05$	.223	.326	.403	.317	.358	.570	.677	.535
	.10	$\Delta_{sj} = +0.10, \Delta_{gj} = -0.10$	.928	.957	.979	.955	.998	1.000	1.000	.999
		$\Delta_{sj} = -0.10, \Delta_{gj} = +0.10$	.647	.763	.860	.757	.866	.960	.990	.939
		$\Delta_{sj} = +0.10, \Delta_{gj} = 0$	.684	.393	.230	.436	.934	.671	.410	.672
		$\Delta_{sj} = 0, \Delta_{gj} = +0.10$	.511	.762	.856	.710	.759	.962	.990	.903
		$\Delta_{sj} = -0.10, \Delta_{gj} = 0$	.510	.279	.173	.321	.769	.454	.267	.497
		$\Delta_{sj} = 0, \Delta_{gj} = -0.10$	.684	.915	.965	.855	.930	.997	1.000	.976

Continued

Table 7
Continued

			Sample Size							
			$N_R = 500, N_F = 500$				$N_R = 1{,}000, N_F = 1{,}000$			
g_{Rj}/s_{Rj}	DIF Size	DIF Type	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
.3	.05	$\Delta_{sj} = +0.05, \Delta_{gj} = -0.05$	.283	.304	.330	.306	.342	.432	.525	.433
		$\Delta_{sj} = -0.05, \Delta_{gj} = +0.05$	.386	.394	.396	.392	.348	.435	.520	.434
		$\Delta_{sj} = +0.05, \Delta_{gj} = 0$	.243	.156	.117	.172	.295	.170	.113	.193
		$\Delta_{sj} = 0, \Delta_{gj} = +0.05$	.305	.325	.359	.330	.289	.409	.489	.396
		$\Delta_{sj} = -0.05, \Delta_{gj} = 0$	.287	.193	.148	.209	.284	.165	.117	.189
		$\Delta_{sj} = 0, \Delta_{gj} = -0.05$	.258	.311	.339	.303	.292	.437	.521	.416
	.10	$\Delta_{sj} = +0.10, \Delta_{gj} = -0.10$	.638	.766	.864	.756	.871	.961	.990	.941
		$\Delta_{sj} = -0.10, \Delta_{gj} = +0.10$	.663	.743	.819	.742	.712	.871	.958	.847
		$\Delta_{sj} = +0.10, \Delta_{gj} = 0$	.480	.273	.166	.306	.708	.389	.232	.443
		$\Delta_{sj} = 0, \Delta_{gj} = +0.10$	.527	.690	.793	.670	.612	.861	.959	.811
		$\Delta_{sj} = -0.10, \Delta_{gj} = 0$	.478	.296	.198	.324	.558	.294	.186	.346
		$\Delta_{sj} = 0, \Delta_{gj} = -0.10$	.499	.739	.853	.697	.707	.947	.988	.881

Note. g_{Rj} = guessing parameter for the reference group; s_{Rj} = slip parameter for the reference group; K_j = number of attributes required for success on the item; $\Delta_{gj} = g_{Fj} - g_{Rj}$; $\Delta_{sj} = s_{Rj} - s_{Fj}$.

Table 8
Power Rates for Detecting Nonuniform DIF based on the Empirical Wald Statistic Distribution

			Sample Size							
			$N_R = 500, N_F = 500$				$N_R = 1{,}000, N_F = 1{,}000$			
g_{Rj}/s_{Rj}	DIF Size	DIF Type	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
.1	.05	$\Delta_{sj} = +0.05, \Delta_{gj} = -0.05$	.683	.738	.756	.726	.951	.968	.976	.965
		$\Delta_{sj} = -0.05, \Delta_{gj} = +0.05$	.438	.504	.546	.496	.758	.810	.846	.805
		$\Delta_{sj} = +0.05, \Delta_{gj} = 0$	.385	.200	.099	.228	.701	.406	.226	.444
		$\Delta_{sj} = 0, \Delta_{gj} = +0.05$	.260	.411	.494	.388	.483	.710	.803	.665
		$\Delta_{sj} = -0.05, \Delta_{gj} = 0$	.261	.151	.104	.172	.483	.266	.157	.302
		$\Delta_{sj} = 0, \Delta_{gj} = -0.05$	.384	.605	.718	.569	.694	.908	.955	.852
.2	.05	$\Delta_{sj} = +0.05, \Delta_{gj} = -0.05$	.262	.334	.380	.325	.506	.629	.696	.610
		$\Delta_{sj} = -0.05, \Delta_{gj} = +0.05$	.213	.265	.313	.264	.369	.485	.571	.475
		$\Delta_{sj} = +0.05, \Delta_{gj} = 0$	.164	.102	.075	.114	.307	.170	.115	.197
		$\Delta_{sj} = 0, \Delta_{gj} = +0.05$	.155	.253	.288	.232	.256	.453	.561	.423
		$\Delta_{sj} = -0.05, \Delta_{gj} = 0$	.156	.104	.084	.115	.254	.150	.100	.168
		$\Delta_{sj} = 0, \Delta_{gj} = -0.05$	.164	.276	.353	.264	.299	.534	.651	.495
	.10	$\Delta_{sj} = +0.10, \Delta_{gj} = -0.10$	.893	.943	.970	.935	.997	.999	.999	.999
		$\Delta_{sj} = -0.10, \Delta_{gj} = +0.10$	.572	.724	.835	.710	.827	.951	.989	.923
		$\Delta_{sj} = +0.10, \Delta_{gj} = 0$	.605	.344	.192	.380	.909	.641	.383	.644
		$\Delta_{sj} = 0, \Delta_{gj} = +0.10$	.431	.719	.829	.660	.708	.955	.988	.883
		$\Delta_{sj} = -0.10, \Delta_{gj} = 0$	.433	.239	.144	.272	.719	.421	.246	.462
		$\Delta_{sj} = 0, \Delta_{gj} = -0.10$	.606	.890	.955	.817	.905	.997	1.000	.967

Continued

Table 8
Continued

			Sample Size							
			$N_R = 500, N_F = 500$				$N_R = 1{,}000, N_F = 1{,}000$			
g_{Rj}/s_{Rj}	DIF Size	DIF Type	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
.3	.05	$\Delta_{sj} = +0.05, \Delta_{gj} = -0.05$	.088	.141	.190	.140	.180	.293	.423	.299
		$\Delta_{sj} = -0.05, \Delta_{gj} = +0.05$	.173	.220	.248	.214	.201	.301	.422	.308
		$\Delta_{sj} = +0.05, \Delta_{gj} = 0$	.075	.057	.051	.061	.150	.093	.069	.104
		$\Delta_{sj} = 0, \Delta_{gj} = +0.05$	.113	.165	.209	.162	.166	.282	.391	.280
		$\Delta_{sj} = -0.05, \Delta_{gj} = 0$	.103	.081	.066	.083	.147	.090	.071	.102
		$\Delta_{sj} = 0, \Delta_{gj} = -0.05$	.079	.150	.194	.141	.153	.304	.418	.292
	.10	$\Delta_{sj} = +0.10, \Delta_{gj} = -0.10$	.333	.561	.737	.544	.727	.914	.981	.874
		$\Delta_{sj} = -0.10, \Delta_{gj} = +0.10$	.425	.582	.700	.569	.579	.799	.931	.770
		$\Delta_{sj} = +0.10, \Delta_{gj} = 0$	.225	.126	.084	.145	.531	.264	.163	.319
		$\Delta_{sj} = 0, \Delta_{gj} = +0.10$	.298	.504	.659	.487	.459	.780	.932	.723
		$\Delta_{sj} = -0.10, \Delta_{gj} = 0$	.246	.152	.107	.168	.399	.195	.130	.241
		$\Delta_{sj} = 0, \Delta_{gj} = -0.10$	.242	.533	.725	.500	.532	.895	.977	.802

Note. g_{Rj} = guessing parameter for the reference group; s_{Rj} = slip parameter for the reference group; K_j = number of attributes required for success on the item; $\Delta_{gj} = g_{Fj} - g_{Rj}$; $\Delta_{sj} = s_{Rj} - s_{Fj}$.

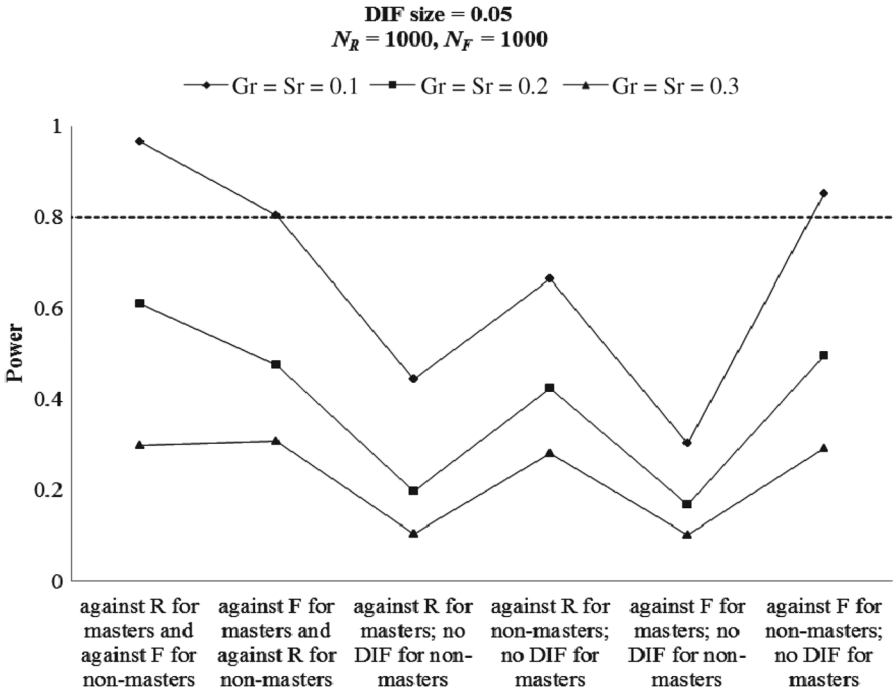

Figure 3. Average empirical power rates by reference item parameter value for nonuniform DIF types with DIF size of .05 ($N_R = N_F = 1{,}000$).

Comparison Study

Previous studies have investigated the performance of MH and SIBTEST when detecting DIF in the context of the DINA model. These two commonly used DIF procedures have been proven to be effective in detecting uniform DIF. In the present study, for the first time, the Wald test was shown to be effective for detecting both uniform and nonuniform DIF. A comparison between the Wald test and the MH and SIBTEST procedures for uniform DIF detection will justify the use of the Wald test over these two traditional methods in the context of the DINA model. To compare the performance of the Wald test and the traditional MH and SIBTEST procedures with the total score as the matching criterion for detecting uniform DIF, a second simulation study was conducted varying the following factors: sample size ($N_R = N_F = 500$ or $N_R = N_F = 1{,}000$), reference item parameters ($g_{Rj} = s_{Rj} = .1; .2; .3$), DIF size ($\Delta_{sj} = \Delta_{gj} = .05; .10$) and proportion of DIF items (10%; 20%; 30%) in the test. Data were generated using a fixed number of attributes ($K = 5$) and fixed test length ($J = 30$) as in the first simulation study. The same Q-matrix was chosen. As before, the joint distributions of attribute profiles were generated with equal probabilities from a multinomial distribution. For each condition, 500 datasets were generated. All three DIF detection methods were applied to each condition. Similar to the first simulation study, the item parameters were estimated separately for the reference and focal group before carrying out the Wald test.

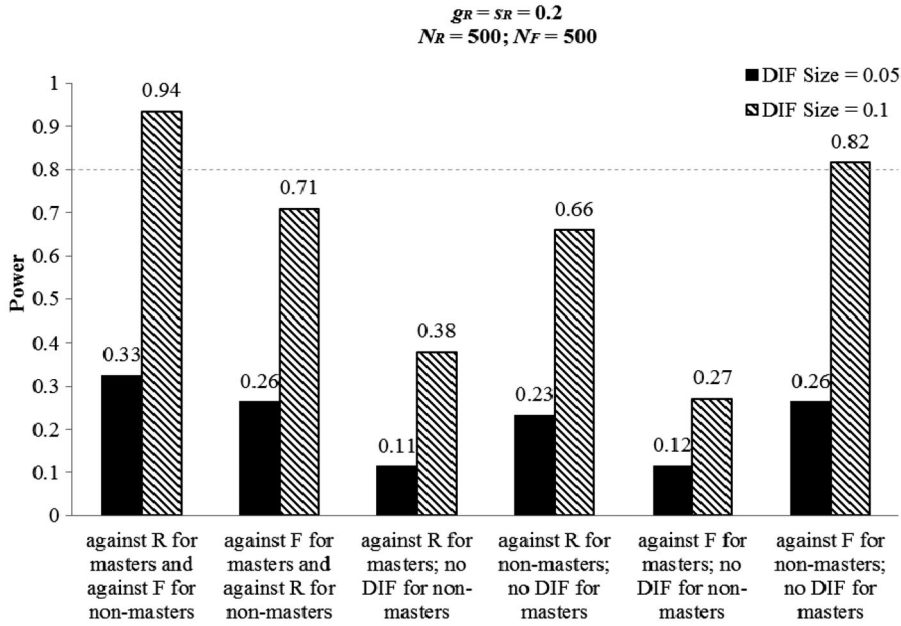

Figure 4. Average empirical power rates for detecting nonuniform DIF by DIF type ($g_R = s_R = .2; N_R = N_F = 500$).

Type I Error Study

Under each condition, the DIF detection rates of the non-DIF items were analyzed for the Type I error rates. Under each condition, the DIF detection rates of the non-DIF items in the test were used to calculate the Type I error rates. The results were similar for the two studied sample sizes. Due to limited space, only the results for the larger sample size are reported. Table 9 summarizes the average Type I error rates comparing the three DIF detection methods when the sample size was 1,000 for both studied groups.

First, it can be seen from Table 9 that the proportion of DIF items in the test did not influence the Type I error rates of the Wald test. This is consistent across the DIF size, DIF type, and reference item parameter values. Comparing the three DIF detection methods, it can be seen that when the item slip and guessing parameters were equal to .1 or .2 the Type I error rates of the Wald test were comparable to those of the MH and SIBTEST procedures if the proportion of DIF items in the test was low (10%); the Type I error rates of the Wald test were closer to the nominal level than those of the MH and SIBTEST procedures if the proportion of DIF items in the test was medium to high (20%–30%). Although the Type I error rates of all three methods were overestimated to some extent, the inflation of the Type I error rates of the MH and SIBTEST procedures was worse than that of the Wald test, especially when the proportion of DIF items in the test increased to 30%. These results also suggest that the traditional MH and SIBTEST procedures are more susceptible to

Table 9
Type I Error Rate Comparison when $N_R = 1{,}000$, $N_F = 1{,}000$

Reference Item Parameter Values	DIF Size	DIF Percentage	DIF Detection Method											
			Wald Test				MH				SIBTEST			
			$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
$g_{Rj} = s_{Rj} = .1$	.05	10%	.064	.051	.055	.057	.056	.053	.053	.054	.054	.055	.054	.055
		20%	.054	.050	.055	.053	.087	.071	.065	.074	.062	.058	.059	.060
		30%	.059	.047	.049	.052	.121	.124	.088	.111	.069	.077	.057	.068
$g_{Rj} = s_{Rj} = .2$	.05	10%	.075	.066	.063	.068	.047	.048	.056	.050	.051	.057	.060	.056
		20%	.078	.061	.070	.070	.068	.070	.079	.072	.067	.072	.081	.073
		30%	.061	.060	.061	.061	.106	.115	.086	.102	.089	.115	.079	.094
	.1	10%	.072	.058	.055	.062	.074	.068	.062	.068	.069	.073	.063	.068
		20%	.075	.058	.056	.063	.167	.163	.131	.153	.131	.149	.108	.129
		30%	.075	.065	.059	.066	.318	.310	.235	.288	.219	.269	.186	.225
$g_{Rj} = s_{Rj} = .3$	.05	10%	.144	.102	.089	.112	.046	.048	.052	.049	.056	.060	.066	.060
		20%	.138	.104	.083	.108	.075	.070	.064	.070	.082	.080	.072	.078
		30%	.137	.098	.082	.106	.101	.102	.089	.097	.104	.119	.104	.109
	.1	10%	.142	.098	.084	.108	.066	.065	.062	.064	.073	.082	.076	.077
		20%	.146	.109	.086	.114	.148	.152	.128	.143	.145	.167	.131	.148
		30%	.139	.108	.092	.113	.269	.287	.242	.266	.232	.301	.229	.254

Note. g_{Rj} = guessing parameter for the reference group; s_{Rj} = slip parameter for the reference group; K_j = number of attributes required for success on the item; $\Delta_{gj} = g_{Fj} - g_{Rj}$; $\Delta_{sj} = s_{Rj} - s_{Fj}$.

Table 10
Power Rate Comparison when $N_R = 1{,}000, N_F = 1{,}000$

Reference Item			DIF Detection Method											
			Wald Test				MH				SIBTEST			
Parameter Values	DIF Size	DIF Percentage	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
$g_{Rj} = s_{Rj} = .1$	.05	10%	.914	.900	.858	.891	.536	.606	.706	.616	.590	.654	.772	.672
		20%	.895	.909	.880	.895	.409	.503	.678	.530	.543	.615	.771	.643
		30%	.892	.901	.895	.896	.310	.400	.599	.436	.481	.565	.761	.602
$g_{Rj} = s_{Rj} = .2$	.05	10%	.662	.640	.624	.642	.574	.558	.586	.573	.594	.572	.592	.586
		20%	.623	.636	.664	.641	.427	.456	.530	.471	.485	.490	.591	.522
		30%	.650	.653	.657	.653	.318	.364	.450	.377	.409	.415	.521	.448
	.1	10%	1.000	1.000	.996	.999	.976	.990	.996	.987	.984	.994	.998	.992
		20%	.997	.999	.997	.998	.944	.963	.977	.961	.966	.972	.985	.974
		30%	.993	.999	.999	.997	.867	.903	.953	.908	.911	.926	.972	.936
$g_{Rj} = s_{Rj} = .3$	.05	10%	.534	.506	.556	.532	.538	.512	.546	.532	.572	.524	.554	.550
		20%	.541	.538	.527	.535	.444	.446	.445	.445	.477	.473	.493	.481
		30%	.547	.519	.532	.533	.360	.343	.364	.356	.400	.375	.411	.396
	.1	10%	.940	.986	.988	.971	.990	.990	.992	.991	.982	.990	.992	.988
		20%	.953	.970	.978	.967	.951	.939	.949	.946	.956	.946	.958	.953
		30%	.952	.969	.983	.968	.867	.855	.902	.874	.895	.871	.919	.895

Note. g_{Rj} = guessing parameter for the reference group; s_{Rj} = slip parameter for the reference group; K_j = number of attributes required for success on the item; $\Delta_{gj} = g_{Fj} - g_{Rj}$; $\Delta_{sj} = s_{Rj} - s_{Fj}$.

contamination of DIF items in the test, whereas the Wald test is more sensitive to the item slip and guessing parameter values.

Power Study

Table 10 provides a summary of power rates for all three DIF detection methods when the sample size is 1,000 for both studied groups. Under each condition, the DIF detection rates of the DIF items in the test were used to calculate the power rates.

It can be seen that when uniform DIF items were present in the test the Wald test was comparable to or outperformed the MH and SIBTEST procedures. The power of the MH and SIBTEST procedures decreased when the proportion of DIF items increased. In contrast, the power of the Wald test was not affected by the proportion of DIF items in the test.

To summarize the results of the comparison study, the proportion of DIF items in the test does not affect the DIF detection rates of the Wald test. This is because the Wald test only requires estimating item parameters separately for the two studied groups—not for the combined sample—so that the existence of DIF items does not distort the item parameter estimates nor the attribute profile estimates for the Wald test. However, the MH and SIBTEST procedures are susceptible to the presence of DIF items in the test. Too many DIF items can contaminate the conditioning variable in the MH and SIBTEST procedures. When there were more DIF items present in the test, the Type I error rates were inflated and the power rates decreased for the MH and SIBTEST procedures. Overall, the Wald test was comparable to or outperformed the MH and SIBTEST procedures when detecting uniform DIF.

Summary and Discussion

CDMs are promising methodologies that can provide more detailed information that is useful in furthering classroom instruction and students' learning. With these psychometric models, novel assessments and data can be generated. However, investing effort and resources to design assessments that can produce diagnostic data is contingent on the assurance that the methodological infrastructure for their analysis and use is firmly in place. The invariance of item parameters should be checked to assure the proper use of CDMs. DIF analysis serves such a purpose. In this use, DIF analysis is critical for test validation to investigate whether the groups identified ahead of time influence test inference. To ensure the validity of the cognitive diagnostic assessment system, developing an efficient method of detecting DIF items is imperative. This study proposes the Wald test for DIF detection in the context of the DINA model. By varying the sample size, reference item parameters, DIF size, and DIF type, the performance of the Wald test to detect DIF was assessed for the DINA model. Results of this study suggest that both for small and large sample sizes (500 and 1,000 examinees for both studied groups) and relatively discriminating items (reference item parameter values less than .3), the Wald test has Type I error rates close to the nominal level. Moreover, its viability is underscored by the medium (around .6) to high (above .8) power rates for most DIF types investigated when DIF size is large (50% of the reference item parameter values).

The second simulation study was conducted to compare the DIF detection performance of the proposed Wald test and the traditional MH and SIBTEST procedures with total test score as the matching criterion. The results showed that the Wald test is a promising approach to detecting DIF in the context of CDMs. Although under some simulation conditions (i.e., when the item parameter values are high and DIF size is small) the Wald test has inflated Type I error and low power rates, its performance was comparable to or outperformed the MH and SIBTEST procedures overall. It should be noted that the MH and SIBTEST procedures were affected by the contamination of DIF items. Additionally, when the proportion of DIF items in the test increased, the performance of these two procedures was further impaired. In contrast, the performance of the Wald test was not affected by the proportion of DIF items in the test.

This study can be viewed as a first step in studying other issues concerning the use of the Wald test to detect DIF in the context of CDMs. It is noted that the proposed method is based on parameter estimation of cognitive diagnostic modeling, which requires a relatively large sample size. In this study it was found that for the sample size of 1,000 in each group the Wald test performed well in detecting DIF when the item is relatively discriminating and DIF size is not too small. The likely testing contexts to which the proposed method can be applied are large-scale educational and psychological assessments. The effectiveness of the Wald test also was explored for small sample sizes for both reference and focal groups. The results showed that the Wald test was effective for detecting both uniform and nonuniform DIF where both masters and nonmasters were involved, even when sample size was small, as long as the item was relatively discriminating and DIF size was large.

Other issues are related to the number of attributes involved in the test and the test length. The simulation design in this study fixed both the number of attributes and the test length. In practice, there are different numbers of attributes involved in an assessment and the tests could be of varying lengths. It will be interesting to investigate whether or not different numbers of attributes or test lengths would affect the performance of the Wald test to detect DIF in the context of CDMs. In addition, the DINA model implemented in this study is one of the most restrictive CDMs. It will be necessary to assess the performance of the Wald test to detect DIF in a more generalized CDM such as the G-DINA model (de la Torre, 2011) and LCDM (Henson, Templin, & Willse, 2009) which allow for various probabilities of success for different attribute profiles. To address these issues thoroughly, applying the Wald test for DIF detection in the context of the G-DINA model incorporating the above discussed factors merits a separate and systematic study. The proposed Wald test provides a way to both improve CDM modeling and also open other possibilities for detecting DIF in complex CDMs.

Finally, we would like to note that DIF can be construed as the presence of a construct-irrelevant latent dimension (possibly more) that is differently distributed across the groups of interest (Shealy & Stout, 1993b). Thus, in identifying this construct-irrelevant dimension and using the corresponding multidimensional models, DIF in unidimensional IRT models can be accounted for. Although CDMs are inherently multidimensional latent variable models, they do not preclude the presence of construct-irrelevant dimensions that do not have the same distributions across

groups. For this reason, as in unidimensional IRT, it would be worthwhile to examine how CDMs with higher dimensions can be used to account for the presence of DIF in the CDM context.

References

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

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115–130.

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

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

de la Torre, J., & Lee, Y. S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement*, *47*, 115–127.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, *39*, 1–38.

Doornik, J.A. (2002). 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*, 333–352.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191–210.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the MH procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19–60). Cambridge, UK: Cambridge University Press.

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.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the MH statistic. *Educational and Psychological Measurement*, *52*, 443–451.

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

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

Narayanon, P., & Swaminanthan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*, 257–274.

Rogers, S. J., & Swaminathan, H. (1993). A comparison of logistic regression and MH procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*, 105–116.

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 Perspective*, *6*, 219–262.

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Shealy, R., & Stout, W. F. (1993a). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–329). Hillsdale, NJ: Lawrence Erlbaum.

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

Swaminathan, H., & Rogers, S. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–370.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*, 251–275.

Templin, J., & Henson, R. A. (2006). Measurement of psychology disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287–305.

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

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.

Authors

LIKUN HOU is Psychometrician, Examinations Team, American Institute of CPAs, Princeton South Corporate Center, 100 Princeton South Suite 200, Ewing, NJ 08628; lhou@aicpa.org. Her primary research interests include item response theory, cognitive diagnostic modeling, differential item functioning, and item response forensic analysis.

JIMMY DE LA TORRE is Associate Professor of Educational Psychology at Rutgers University, 10 Seminary Place, New Brunswick, NJ 08901; j.delatorre@rutgers.edu. His primary research interests include item response theory, cognitive diagnosis modeling, Bayesian analysis, and the use of diagnostic assessments to support classroom instruction and learning.

RATNA NANDAKUMAR is Professor in the School of Education at the University of Delaware, 213 Willard Hall Education Building, Newark, DE 19716; nandakum@udel.edu. Her primary research interests include applications of item response theory, dimensionality, and DIF.