EXAMINING EXISTING METHODS OF BIAS DETECTION IN STANDARDIZED TESTING

BY

EMMA THRONSON

HONORS THESIS

Submitted in partial fulfillment of the requirements
for the degree of Bachelor of Science (Honors Data Science)
from the College of Literature, Science, and the Arts at the
University of Michigan, April 2023

Ann Arbor, Michigan

Advisor:

Gongjun Xu

Acknowledgements

I am grateful for the valuable guidance and support provided by Professor Gongjun Xu and Jing Ouyang throughout the research process and the writing of this paper over the last year. Their insightful feedback and advice have greatly enhanced this work. I am also grateful to the Statistics Department for providing me the opportunity to become involved in research via the Undergraduate Research Program in Statistics, and for providing me the necessary resources and facilities to carry out this study.

Abstract

The detection of differential item functioning (DIF), or bias, in standardized testing is critical to ensure the validity and fairness of test results. Ensuring equity and fairness to all demographic subgroups is critical for these exams. Cognitive Diagnosis Models (CDMs) may be generated from test data to provide examinees with feedback on skill mastery. This study examined the effectiveness of existing methods of DIF-identification in CDM models generated from standardized test data. The Wald test, the Likelihood Ratio (LR) test, the Benjamini-Hochberg (BH) Procedure, the Mantel-Haenszel (MH) test, and the McNemar test were all implemented to identify DIF in DINA (deterministic inputs, noisy "and" gate) models, a type of CDM. Performance changes based on sample size, test length, and number of attributes were examined by simulating data; these methods were also examined using real-world data from the TIMSS 2007 and PISA 2000 exams. The identification methods were evaluated, compared, and suggestions were made to users on how to best utilize these existing tools for DIF-identification.

1 Introduction

Differential item functioning (DIF) detection in a standardized test is an important task to ensure the validity and fairness of the results. DIF analysis provides an assessment of measurement invariance. Measurement invariance is an important property in "establishing the fairness and construct validity" of assessments such as standardized tests and psychological assessments; i.e., presence of invariance indicates the test is unbiased [Mehrazmay et al., 2021]. Ideally, an examinee's ability to successfully answer a test item should rely solely on their underlying skills, such as the ability to reduce fractions or identify the subject in a sentence; that is, any differences in performance between examinees should be explained wholly by their mastery of these skills [Mehrazmay et al., 2021]. Detecting DIF in a particular item from a standardized test indicates that the response distribution for that item relies on subgroup membership, such as gender, race, or age, in addition to the aforementioned underlying latent traits. Presence of DIF indicates there is bias present in the item, whereas an unbiased item response would rely solely on the latent attributes, or "attribute profile" of the examinee [de la Torre, 2011]. An item displays DIF if at least two subgroups have varying probabilities of success on a particular item even when the examinees' underlying latent attributes are the same. Standardized assessments aim to be measurement invariant, or DIF-free, to be fair to all examinees regardless of subgroup membership. Proper DIF analysis can help test-givers determine whether this goal has been achieved. This thesis will evaluate the effectiveness of existing methods to properly identify differential item functioning (bias) in standardized testing.

Cognitive Diagnosis Models (CDMs) are one example of an instrument in which DIF can be detected. CDMs provide feedback to examinees about the attributes they have or have not mastered [Mehrazmay et al., 2021]. Therefore, it is important to know if any items in the CDM display DIF to ensure assessment results are not skewed by subgroup membership. CDMs are explained in greater detail in Section 2.1.

There are multiple kinds of CDMs, including DINA (deterministic inputs, noisy "and" gate), DINO (deterministic inputs, and noisy "or" gate, and ACDM (additive CDM). All of these are special cases of the generalized DINA (GDINA) model [Ma and de la Torre, 2020]. DIF-detection methods exist for all of these models. In this study, the performances of some commonly used DIF-detection methods are demonstrated and compared for DINA models. These methods include the Wald test, Likelihood Ratio (LR) test, Benjamini-Hochberg (BH) procedure, Mantel-Haenszel (MH) test, and McNemar test. These are implemented using multiple R packages including the `CDM` package authored by Robitzsch et al. and the `GDINA` package authored by Ma and de la Torre [Ma and de la Torre, 2020, Robitzsch et al., 2020], and a variety of built-in functions [R Core Team, 2021].

These existing tests for assessment leave room for improvement due to lack of generalization, meaning they are only effective under certain settings, e.g. larger sample sizes or fewer latent attributes. In order to understand which aspects of the existing DIF-detection methods require improvement, it is critical to have a robust understanding of where they fail. In this paper, existing methods of DIF-detection, specifically the Wald and LR tests executed by the `CDM` and `GDINA` R packages mentioned previously, are evaluated under a number of settings, such as varying sample size, length of the exam, and number of attributes. A previous study conducted by Hou et al. conducted the Wald test while varying sample size to compare how performance changes, however this study did not investigate varying test lengths or number of attributes [Hou et al., 2014]. This study will conclude with a discussion of when these methods fail under certain conditions. Additionally, this paper identifies where improvements can be made and offers recommendations for users attempting to detect DIF. This problem of DIF-detection warrants close attention and improvement to ensure increased fairness in standardized testing.

This thesis evaluates and compares existing statistical tests often used for the purpose of identifying DIF in standardized testing.

2 Models and Methods

2.1 Cognitive Diagnosis Models

Cognitive Diagnosis Models (CDMs) are powerful tools that can provide feedback about an examinee's mastery of different latent attributes based on their test results [Mehrazmay et al., 2021]. DIF identification can be performed on a variety of CDMs. There are different benefits to creating and using each type of model, however this study limits its examination to the DINA (**d**eterministic **i**nputs, **n**oisy "**a**nd" gate) model which provides greater interpretability.

As detailed by de la Torre, the Q matrix is a binary matrix of dimensions $J \times K$, where J is the number of items on the test and K is the number of latent attributes this test examines [de la Torre, 2011]. For each row (each item), the cell q_{jk} is 1 if that latent attribute is necessary to answer the test item correctly, and 0 otherwise. The data, or item response, matrix X is a binary matrix of dimensions $N \times J$, where N is the number of examinees. Each row (examinee) represents the success vector of an individual test-taker, with a 1 in the cell if they answered the item correctly and a 0 otherwise. Finally, each examinee has a latent attribute vector α of dimensions $1 \times K$, which represents the underlying skills and attributes that particular test taker possesses. Once again, a 1 represents possession of the attribute while a 0 indicates otherwise.

Additionally, guessing and slipping parameters are defined. The guessing parameter represents the probability that an examinee who does not possess the necessary attributes for an item could correctly guess on that item. The slipping parameter represents the probability that an examinee who does possess the necessary attributes for an item could "slip" and get the item wrong [de la Torre, 2011].

To represent DIF in CDMs, the following value is computed:

$$\Delta_{j\alpha_i} = P(X_j = 1|\alpha_i)_F - P(X_j = 1|\alpha_i)_R$$

where F represents the focal group, R represents the reference group, and $\Delta_{j\alpha_i}$ represents DIF in item j when an examinee possesses the latent attribute vector α_i [Hou et al., 2014]. This $\Delta_{j\alpha_i}$ represents the difference in success probability for an examinee in the focal group versus an examinee in the reference group when they have the same attribute possession. A negative value indicates the reference group has a higher probability of success, while a positive value indicates the focal group has a higher probability of success.

2.1.1 DINA

There are many types of CDM which make differing assumptions about the data. The CDM used in this study was the DINA model. The DINA (**d**eterministic **i**nputs, **n**oisy "**a**nd" gate) model is best applied when all required attributes for a test item must be mastered or present in order for the examinee to answer the item correctly [Hou et al., 2014]. With this model, lacking one or more attribute necessary for an item is equivalent to lacking them all. The DINA model was selected for this project due to this simplifying assumption and ease of interpretation. With the DINA model, the probability of an examinee correctly answering a test item is as follows:

$$P_j(\alpha_i) = P(X_{ij}|\alpha_i) = (1 - s_j)^{\eta_{ij}} {g_j}^{1-\eta_{ij}}$$

where α_i is the latent attribute vector associated with examinee i, s_j and g_j are the slipping and guessing parameters associated with item j, and η_{ij} is as follows:

$$\eta_{ij} = \prod_k^K \alpha_{ik}^{q_{jk}}.$$

$\eta_{ij} = 1$ when the examinee has all attributes necessary to succeed on item j, and $\eta_{ij} = 0$ when they are missing at least one of these required attributes [Hou et al., 2014]. As detailed further by Hou, DIF in an item is detected in a DINA model if the guessing and/or slipping parameters of that item for one group differ from those of the other group.

To test the accuracy and effectiveness of existing models, this study focused on the Wald and LR tests to identify items with DIF, performed by the `CDM` and `GDINA` packages in R. The packages were first used to construct a DINA CDM model (object) when provided with a Q matrix, attributes, and results data from a real-world or simulated test. Then, they took these models and ran statistical tests, Wald or LR, returning the likelihood of a particular item displaying DIF. In particular, a test statistic and a p-value were calculated for each test item; a significant p-value ($\alpha = 0.05$) indicated the item displays DIF.

2.2 Statistical Tests

A variety of statistical tests were performed on the simulated data and the real-world data in this study.

2.2.1 Wald Test

The Wald test can be applied via multivariate hypothesis testing, examining if the slipping and/or guessing parameters are different between different subgroups. The null hypothesis of the Wald test is that there is no difference in the guessing or slipping parameters between groups. To perform the Wald test, an unconstrained model is fit to the data. Item parameters are estimated as follows:

$$\hat{\beta}_j^* = (\beta_{Rj}, \beta_{Fj}) = (g_{Rj}, s_{Rj}, g_{Fj}, s_{Fj})'.$$

Next, a constrained model is fit, meaning item parameters are constrained to be equal across subgroups. The fit of both models is compared to see if constraining the model reduces the fit to the data [Hou et al., 2014]. The Wald test statistic follows a χ^2 distribution with K degrees of freedom. It is computed as follows:

$$W_j = [R_j \times \hat{\beta}_j^*]'[R_j \times \mathrm{Var}(\hat{\beta}_j^*) \times R_j']^{-1}[R_j \times \hat{\beta}_j^*]$$

where R_j is a restriction matrix and $\mathrm{Var}(\hat{\beta}_j^*)$ is the variance-covariance matrix of item parameters [Mehrazmay et al., 2021, Hou et al., 2014]. These elements are implemented as

$$R_j = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}_{2\times 4} \quad \text{and} \quad \mathrm{Var}(\hat{\beta}_j^*) = \begin{bmatrix} \mathrm{Var}(\hat{\beta}_{Rj}^*) & 0 \\ 0 & \mathrm{Var}(\hat{\beta}_{Fj}^*) \end{bmatrix}_{2\times 2}.$$

The function `CDM::gdina.dif` makes its decisions using the Wald test. The function `GDINA::dif` also makes its decisions using the Wald test if this is specified as the "method" argument of the function. In the `CDM` package, the Wald test is applied to each item individually, while in the `GDINA` implementation, the Wald test is applied to all items simultaneously. Both packages compare the Wald test statistics to a standard normal distribution in order to determine p-values. The use and details of the Wald test in a CDM application are described in greater detail in [Mehrazmay et al., 2021].

2.2.2 Likelihood Ratio Test

The Likelihood Ratio (LR) test similarly compares two models: a reduced model and an augmented model [Mehrazmay et al., 2021]. The LR test statistic is computed as follows:

$$G^2 = -2[LL_{\text{reduced}} - LL_{\text{augmented}}]$$

where LL represents the log-likelihoods of the specific model. This test statistic follows a χ^2 distribution. The degrees of freedom for this statistic is the difference between the number of parameters of the reduced and augmented models [Ma et al., 2021]. The null hypothesis of this test is that no DIF is present. The function `GDINA::dif` can also make its decisions using the LR test if this is specified as the "method" argument of the function. The `CDM` package does not have this test option for DIF-identification. The use and details of the LR test in a CDM application are described in greater detail in [Mehrazmay et al., 2021].

2.2.3 Benjamini-Hochberg Procedure

The Benjamini-Hochberg (BH) Procedure introduces the concept of controlling for false positives. False positives in the context of this research indicate a test item that was incorrectly said to display DIF. This is due to the fact that sometimes significant p-values occur by chance. In this study, the False Positive Rate (FPR) was shown for the Wald and LR tests in the tables present in Sections 3.1.2 and 3.1.3. Applying the BH procedure allows for determining if a statistical test, the Wald test in this case, has over-identified DIF; i.e., has a a false-discovery rate higher than 10%.

This procedure requires taking existing p-values produced by the model and creating a new critical value to determine which significant items may have been false positives. The selected false positive rate for this data was 10%, however another typical value taken is 5%. To extract p-values, a DINA model was created from the gender-categorized data, and the CDM DIF-identification function was applied. The Wald test results from the `CDM` DIF function were selected over the `GDINA` DIF function due to greater robustness of the Wald test provided by `CDM` package to a wider variety of settings, as discussed later. After p-values were extracted, then they were arranged in increasing order and a rank was assigned to each item, 1 through J where J is the number of items in the data. Next, a BH critical value was calculated for each test item using the following formula:

$$\text{critical value} = \frac{\text{item rank}}{J} \times q$$

where q is the specified false positive rate (10% in this study). Finally, the item with the highest p-value that is still lower than its corresponding critical value was marked. That item and any item above it were considered to be statistically significant, and were the items identified as displaying DIF with a false discovery rate of 10%.

2.2.4 Mantel-Haenszel Test

The Mantel-Haenszel (MH) test is a useful baseline metric for identifying DIF in real-world data, where the true labels of DIF are unknown. This test is based on contingency tables constructed from the counts of the examinees' success on each test item. The null hypothesis of this test is that the success ratio of both groups are the same [Wainer and Sireci, 2005].

The data collected from the `CDM` package provided access to the number of examinees who successfully answered each test items, and the number who did not. With this information, it was possible to create 2×2 contingency tables where each row corresponded to a group (male or female) and each column corresponded to the success of said group (correct or incorrect). To take this one step further, strata were introduced to further group the data. In this scenario, the strata were the test items. A $2 \times 2 \times K$ table was then be constructed, where K is the number of items associated with one exam. Finally, this table was passed into the R function `mantelhaen.test` from the `stats` package to generate the test statistic [R Core Team, 2021].

A χ^2 test statistic with one degree of freedom is calculated and a p-value is provided based on this statistic. The null hypothesis of the MH test is that common odds ratio is 1, meaning there is not significant association between success and group membership within each stratum (test item). If the p-value is significant, i.e. below the threshold $\alpha = 0.05$, then according to this test, there is in fact association between success and group identity within each stratum. In this scenario, this indicates that gender does in fact play a role in an examinee's success on a test item. This result is generalized for the whole test at one time.

The MH test is based entirely on the gender subgroup each examinee belongs in, and simple counts of how well that group performed on each test item. Unlike the BH procedure (see Section 2.2.3), the p-values obtained from a DIF-identifying function such as Wald or LR are not used. Only the counts extracted from the X (item success) matrix provided with the data were required for this test.

This MH test is a straightforward baseline for the purpose of this research. It does not account for whether the skills of the examinees are matched and relies solely on the counts of how many examinees from each group answered particular items correctly. This test takes into account all test items at one time, so it is necessary to introduce the McNemar test for item-specific evaluation.

2.2.5 McNemar Test

McNemar's test is almost identical to the MH test, save that it allows an input of a single 2×2 contingency table (a single test item's results) at a time, rather than a stratified $2 \times 2 \times K$ table. These contingency tables are the same as those constructed for the MH test, and an example can be seen in Table 8 or Table 15. This individual contingency table for each single item is provided as input to the `mcnemar.test` function from the `stats` package to generate the test statistic [R Core Team, 2021]. The resulting χ^2 test statistic and p-value are associated with this single item. A significant p-value indicates that membership in a particular gender subgroup does affect the outcome of whether an examinee is successful on that particular test item.

3 Data Analysis and Results

3.1 Simulated Data

3.1.1 Creating Simulating Data

The first part of this study focused on examining the success that the Wald and LR tests have on DIF-identification of simulated data, as provided by the `CDM` and `GDINA` packages. The simulated data was created in part using functions available within these packages. Due to the nature of these implementations, a DINA model must be created from simulated data, meaning the components of a CDM must be simulated. These elements of a CDM are described in Section 2.1.

First, the number of examinees (N), test items (J), latent attributes (K), and guessing and slipping parameters were specified. In this study, N varied from $[500, 1000]$, J varied from $[20, 40, 60]$ and K varied from $[2, 5, 9]$. The guessing and slipping parameters for each item were small decimals, specified for each subgroup.

A Q matrix of size $J \times K$ was a hand-constructed binary matrix indicating which attributes were necessary for success on each item.

$$Q = \begin{bmatrix} 1 & 0 & \cdots & 1 & 1 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{J \times K}$$

A vector of size $1 \times N$ was constructed which indicates in a binary fashion which subgroup each examinee belongs to.

$$\text{groupings} = \begin{bmatrix} 1 & 0 & \cdots & 1 & 1 \end{bmatrix}_{1 \times N}$$

In these simulated settings, these subgroups were of equal size. Guessing and slipping parameters were defined for each test item for the first (control) group. For the second (treatment) group, one fifth of the items' guessing and slipping parameters were modified to differ from those of the first group. This introduced bias into the simulated data, making one subgroup more or less likely to perform well on these modified items.

$$\text{guessing}_{\text{group 1}} = \begin{bmatrix} 0.2 & 0.2 & \cdots & 0.2 & 0.2 \end{bmatrix}_{1 \times J}$$

$$\text{guessing}_{\text{group 2}} = \begin{bmatrix} 0.3 & 0.2 & \cdots & 0.3 & 0.2 \end{bmatrix}_{1 \times J}$$

In the example above, the second and second to last items were intentionally modified to be biased for the second group. The IDs of these biased items were stored in order to determine if the Wald and LR tests ultimately identified them successfully. The accuracy of these tests was determined by how often they could correctly identify these modified items. Finally, after the Q matrix, group assignments, and guessing and slipping parameters were defined, they were passed into the `CDM::sim.din` function from the `CDM` package in order to produce a simulated data (X) matrix which contained the simulated responses of each examinee for each test item.

$$X = \begin{bmatrix} 1 & 1 & \cdots & 1 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & 1 \end{bmatrix}_{N \times J}$$

Then, the `CDM::gdina` function was used to construct a DINA model based on the simulated Q matrix, X matrix, and group identification vector. This DINA model was passed to the `CDM::gdina.dif`, which produced an object (which is referred to as a DIF object) containing the test statistic and p-values associated with each test item.

The `GDINA` package has equivalent functions as `CDM` for the task of simulating data. The `GDINA::simGDINA` function takes a Q matrix, group assignments, and guessing and slipping parameters in order to simulate the X matrix. This function also encodes the type of CDM, DINA in this study, to be used. Because this model type is identified in the data creation phase, there was no need for an intermediate step of creating a DINA model. The `GDINA::dif` took the Q matrix, the X matrix, and the group assignments to produce the DIF object. This `GDINA::dif` function additionally took in a parameter to specify the statistical test to be used, either Wald or LR.

The DIF object created by either the `CDM` and `GDINA` packages was the source of determining if the identification of bias was successful. From this DIF object, a p-value was extracted from these test statistics for each item. Ultimately if the p-value was significant (less than or equal to the threshold $\alpha = 0.05$), then the associated item had been flagged as displaying DIF. These identified items were then cross-checked with the stored list of items that were intentionally modified in order to see if the R package was able to flag these specific items.

To test the Wald and LR test DIF-identification methods, this paper simulated standardized testing data for two examinee subgroups, one with and one without introduced bias. This subgroup is ambiguous, but could represent age, gender, race, or any other demographic category. In order to examine the robustness of the Wald and LR tests, a variety of size settings were used. This was accomplished by varying the number of examinees, number of test items, and number of latent attributes each examinee might possess as described earlier in this section. From these simulated settings, the `CDM` and `GDINA` packages were used to create the Q matrix, data matrix, and latent attribute matrix, which were then used to create CDM objects (in this case, a DINA object) for each combination of these settings. This means there were 18 different simulated models to examine for the Wald test from `CDM`, 18 for the Wald test from `GDINA`, and 18 for the LR test from `GDINA`. For each setting, 1000 simulations were run.

Using simulated data allowed for a more exact analysis of how well the methods perform, as the intentionally biased items were tracked and compared to the test results to see if the methods could pick up on these items. The Wald test was implemented in two manners as described in Section 3.1.2, while the LR test was implemented as described in Section 3.1.3. The tables associated with these implementations are presented below for ease of reference.

3.1.2 Wald Test

The Wald test, as described in Section 2.2.1, is a common method by which DIF is identified in CDMs. There are two manners in which the Wald test can be performed in R: the `CDM` package and the `GDINA` package.

First, the `CDM` package was used. A Q matrix, attributes, X matrix, and group divisions were all generated using the `CDM` package. From these simulated factors, a DINA object was constructed. Next, the DINA model was passed into the `CDM::gdina.dif` function. This function, from the `CDM` R package, "assesses item-wise differential item functioning in the GDINA model by using the Wald test" [Robitzsch et al., 2020]. The `CDM` package's DIF-identifying function uses the Wald test to produce its test statistics and p-values. This process of simulating data is described in greater detail in Section 3.1.1.

Table 1 and Table 2 display the results of the Wald test as conducted by the `CDM` package. As an example, take the setting N = 500, I = 20, K = 2. Since 1000 simulations were run for this single setting, there were a total of 20,000 items, with 4,000 (one fifth of the items) being "Underlying Positives" (with DIF) and 16,000 being "Underlying Negatives" (no DIF). "True Positives" indicates the number of underlying positives that the Wald test accurately identified with DIF. "False Negatives" indicates the number of underlying positives which the Wald test failed to identify as displaying bias. This is a Type II error. "True Negatives" indicates the number of underlying negatives that the Wald test accurately found to be unbiased. "False Positives" indicates the number of underlying negatives that the Wald test incorrectly identified as displaying bias. This is a Type I error. In the conditions of this study, a Type II error was considered more serious; false negatives, or failure to flag a biased question, were the most harmful in this scenario, as this would allow biased items to continue to be used, causing an unfair environment for test takers. The "True Positive Rate" (TPR) was calculated by taking the proportion of true positives out of all items of the underlying positive group ($\frac{TP}{TP+FN}$). Similarly, the "False Negative Rate" (FPR), "True Negative Rate" (TNR), and "False Positive Rate" (FPR) were all calculated from the proportions of true positives, true negatives, false positives, and false negatives. One instance of a simulation was considered "correct" if the test accurately identified the specific items which were modified to display bias, and only these items. The final row in the results table indicates how many of the 1000 simulations for each setting were successful in this task.

The Wald Test could also be conducted via the `GDINA` package. A Q matrix, attributes, X matrix, and group divisions were all generated using the `GDINA` package. These simulated factors were then run through the `GDINA::dif` function, with the model type specified as DINA and the method specified as Wald. This process of simulating data is described in greater detail in Section 3.1.1.

The settings for this set of simulations followed the same pattern as above, and the results tables (Table 3 and Table 4) additionally followed the same format.

3.1.3 Likelihood Ratio Test

The Likelihood Ratio (LR) test, as described in further detail in Section 2.2.2, is another common method of DIF-identification. This may also be conducted via the `GDINA` package.

A Q matrix, attributes, X matrix, and group divisions were all generated using the `GDINA` package. These simulated factors were then run through the `GDINA::dif` function, with the model type specified as DINA and the method specified as LR. This process of simulating data is described in greater detail in Section 3.1.1.

The settings for this set of simulations followed the same pattern as in Section 3.1.2, and the results tables (Table 5 and Table 6) additionally followed the same format.

Wald Test, CDM									
N = 500									
	J = 20			J = 40			J = 60		
	K = 2	K = 5	K = 9	K = 2	K = 5	K = 9	K = 2	K = 5	K = 9
Underlying Positives	4000	4000	4000	8000	8000	8000	12000	12000	12000
True Positives	3189	2878	2195	6416	6230	4852	9575	9432	7376
False Negatives	811	1122	1805	1584	1770	3148	2425	2568	4624
Underlying Negatives	16000	16000	16000	32000	32000	32000	48000	48000	48000
True Negatives	15187	15323	15462	30393	30269	29644	45539	45292	44238
False Positives	813	677	538	1607	1731	2356	2461	2708	3762
True Positive Rate	79.73%	71.95%	54.88%	80.20%	77.88%	60.65%	79.79%	78.60%	61.47%
False Negative Rate	20.48%	28.05%	45.13%	19.80%	22.13%	39.35%	20.21%	21.40%	38.53%
True Negative Rate	94.92%	95.77%	96.64%	94.98%	94.59%	92.64%	94.87%	94.36%	92.16%
False Positive Rate	5.08%	4.23%	3.36%	5.02%	5.42%	7.36%	5.13%	5.64%	7.84%
Correct simulations (out of 1000)	370	264	32	133	120	9	56	57	2

Table 1: Simulated results for `CDM` package, N = 500.

Wald Test, `CDM`									
N = 1000									
	J = 20			J = 40			J = 60		
	K = 2	K = 5	K = 9	K = 2	K = 5	K = 9	K = 2	K = 5	K = 9
Underlying Positives	4000	4000	4000	8000	8000	8000	12000	12000	12000
True Positives	3771	3766	2632	7600	7730	5842	11456	11582	88996
False Negatives	229	234	1368	400	270	2158	544	418	3101
Underlying Negatives	16000	16000	16000	32000	32000	32000	48000	48000	48000
True Negatives	15266	15343	15440	30463	30372	29956	45609	45456	44744
False Positives	734	657	560	1537	1628	2044	2391	2544	3256
True Positive Rate	94.28%	94.15%	65.80%	95.00%	96.63%	73.03%	95.47%	96.52%	74.16%
False Negative Rate	5.73%	5.85%	34.20%	5.00%	3.38%	26.98%	4.53%	3.48%	25.84%
True Negative Rate	95.41%	95.89%	96.50%	95.20%	94.91%	93.61%	95.02%	94.70%	93.22%
False Positive Rate	4.59%	4.11%	3.50%	4.80%	5.09%	6.39%	4.98%	5.30%	6.78%
Correct simulations (out of 1000)	791	779	100	666	757	42	569	654	18

Table 2: Simulated results for `CDM` package, N = 1000.

Wald Test, `GDINA`									
N = 500									
	J = 20			J = 40			J = 60		
	K = 2	K = 5	K = 9	K = 2	K = 5	K = 9	K = 2	K = 5	K = 9
Underlying Positives	4000	4000	4000	8000	8000	8000	12000	12000	12000
True Positives	2889	1229	157	5523	1753	4401	7381	3811	10993
False Negatives	1111	2771	3843	2477	6247	3599	4619	8189	1007
Underlying Negatives	16000	16000	16000	32000	32000	32000	48000	48000	48000
True Negatives	15417	15517	15534	31285	31748	15436	47461	42781	6111
False Positives	583	483	466	715	252	16564	539	5219	41889
True Positive Rate	72.23%	30.73%	3.93%	69.04%	21.91%	55.01%	61.51%	31.76%	91.61%
False Negative Rate	27.78%	69.28%	96.08%	30.96%	78.09%	44.99%	38.49%	68.24%	8.39%
True Negative Rate	96.36%	96.98%	97.09%	97.77%	99.21%	48.24%	98.88%	89.13%	12.73%
False Positive Rate	3.64%	3.02%	2.91%	2.23%	0.79%	51.76%	1.12%	10.87%	87.27%
Correct simulations (out of 1000)	249	3	1	41	0	114	1	0	563

Table 3: Simulated results for `GDINA` package, Type = Wald, N = 500.

Wald Test, `GDINA`									
N = 1000									
	J = 20			J = 40			J = 60		
	K = 2	K = 5	K = 9	K = 2	K = 5	K = 9	K = 2	K = 5	K = 9
Underlying Positives	4000	4000	4000	8000	8000	8000	12000	12000	12000
True Positives	3824	2606	470	5785	5380	4655	11293	7411	10935
False Negatives	176	1394	3530	415	2620	3345	707	4589	1065
Underlying Negatives	16000	16000	16000	32000	32000	32000	48000	48000	48000
True Negatives	15293	14383	15656	30885	28449	16827	46668	45094	7742
False Positives	707	1617	344	1115	3551	15173	1332	2906	40258
True Positive Rate	95.60%	65.15%	11.75%	72.31%	67.25%	58.19%	94.11%	61.76%	91.13%
False Negative Rate	4.40%	34.85%	88.25%	5.19%	32.75%	41.81%	5.89%	38.24%	8.88%
True Negative Rate	95.58%	89.89%	97.85%	96.52%	88.90%	52.58%	97.23%	93.95%	16.13%
False Positive Rate	4.42%	10.11%	2.15%	3.48%	11.10%	47.42%	2.78%	6.05%	83.87%
Correct simulations (out of 1000)	839	150	0	646	32	131	478	3	440

Table 4: Simulated results for `GDINA` package, Type = Wald, N = 1000.

Likelihood Ratio Test									
N = 500									
	J = 20			J = 40			J = 60		
	K = 2	K = 5	K = 9	K = 2	K = 5	K = 9	K = 2	K = 5	K = 9
Underlying Positives	4000	4000	4000	8000	8000	8000	12000	12000	12000
True Positives	3106	1746	193	6194	3886	224	9379	5690	200
False Negatives	894	2254	3807	1806	4114	7776	2621	6310	11800
Underlying Negatives	16000	16000	16000	32000	32000	32000	48000	48000	48000
True Negatives	15126	14438	15833	30369	26883	31995	45540	40760	48000
False Positives	874	1562	167	1631	5117	5	2460	7240	0
True Positive Rate	77.65%	43.65%	4.83%	77.43%	48.58%	2.80%	78.16%	47.42%	1.67%
False Negative Rate	22.35%	56.35%	95.18%	22.58%	51.43%	97.20%	21.84%	52.58%	98.33%
True Negative Rate	94.54%	90.24%	98.96%	94.90%	84.01%	99.98%	94.88%	84.92%	100%
False Positive Rate	5.46%	9.76%	1.04%	5.10%	15.99%	0.02%	5.13%	15.08%	0%
Correct simulations (out of 1000)	348	30	0	117	5	0	40	1	0

Table 5: Simulated results for `GDINA` package, Type = LR, N = 500.

Likelihood Ratio Test									
N = 1000									
	J = 20			J = 40			J = 60		
	K = 2	K = 5	K = 9	K = 2	K = 5	K = 9	K = 2	K = 5	K = 9
Underlying Positives	4000	400 0	4000	8000	8000	8000	12000	12000	12000
True Positives	3811	2785	580	7685	6435	1028	11491	9037	1473
False Negatives	189	1215	3420	315	1565	6972	509	2963	10527
Underlying Negatives	16000	16000	16000	32000	32000	32000	48000	48000	48000
True Negatives	15205	13609	15859	30392	20723	31993	45600	34569	48000
False Positives	795	2391	141	1608	11277	7	2400	13431	0
True Positive Rate	95.28%	69.63%	14.50%	96.06%	80.44%	12.85%	95.76%	75.31%	12.28%
False Negative Rate	4.73%	30.38%	85.50%	3.94%	19.56%	87.15%	4.24%	24.69%	87.73%
True Negative Rate	4.97%	14.94%	0.88%	5.03%	35.24%	0.02%	5.00%	27.98%	0%
False Positive Rate	95.03%	85.06%	99.12%	94.98%	64.76%	99.98%	95.00%	72.02%	100.00%
Correct simulations (out of 1000)	824	197	0	730	196	0	581	50	0

Table 6: Simulated results for `GDINA` package, Type = LR, N = 1000.

3.1.4 Comparisons

To compare the capabilities of the Wald and LR tests, the false positive rates (FPR) and false negative rates (FNR) were examined. The FNR was critical to examine as this was the value which determines if these methods allowed biased items to go unidentified. The FPR was a secondary, however still important, value as it indicated whether the methods marked items as biased when they truly were not; this was also important, as a test that identifies every item as biased is no more useful than a test that fails to identify any items as biased.

Between the two methods of performing the Wald test, the `CDM` package appeared to perform much better than the `GDINA` package. Specifically, the average FNR when $N = 500$ was 28.24% versus 51.37%. When $N = 1000$, this value was 12.78% versus 28.91%. Due to this statistic, the Wald test as performed by the `CDM` package was the chosen method of executing the Wald test moving forward in this study.

Between the Wald test and the LR test, when $N = 500$, the FNR of the former was 28.34% while the latter was 57.54%. When $N = 1000$, the FNR was 12.78% versus 38.66%. The Wald test performed much better than the LR test in regards to the FNR.

As can be seen from these average FNRs, increasing the number of examinees affected the success of the Wald test and LR test in a positive way; the FNR decreased when N increased. The best average FNR achieved was 12.87% from the Wald test when $N = 1000$.

Similarly, the FPRs could be compared. Just like with the FNR, the Wald test provided by the `CDM` package had better results that those of the `GDINA` package and was be used as the comparison to the LR test. When $N = 500$, the Wald test produced an FPR of 8.18% while the LR test produced a comparable 6.40%. When $N = 1000$, however, the Wald test produced 5.06% while the LR test produces an average 89.55% FPR. This is very significant, and shows that the LR test overidentified DIF when the sample size was large.

In addition to these rates, the overall accuracy of these tests could be compared by observing the percentage of correct simulations each test produced. When $N = 500$, the Wald test (`CDM` package) had an average accuracy of 11.5% and the LR test had an average accuracy of 6.01%. When $N = 1000$, the Wald test had an average accuracy of 48.6% while the LR test had an average accuracy of 28.64%. Once again, this shows the Wald test outperformed the LR test, and both became more accurate as the number of examinees rose.

Beyond taking these average rates and accuracies when grouped by N, the performance of the Wald and LR tests could be compared as J or K rise. As J increased from 20 to 40 to 60, the accuracy of both the Wald and LR tests decreased. Similarly, as K increased from 2 to 5 to 9, the accuracy of the Wald and LR tests generally decreased. This means that the tests performed the best when K was small, J was small, and N was large.

These trends in accuracy were also reflected in the FNRs and FPRs of these tests. For the Wald test, as J increased, the FNR tended to decrease slightly; as K increased the FNR increased. For the LR test, as J increased there was not much change in the FNR; as K increased, the FPR first rose between $K = 2$ and $K = 5$, and then fell when $K = 9$.

It was clear that the Wald test had more success detecting DIF than the LR test under any given setting, particularly in the accuracy and FNR. It was also evident that a larger N and smaller J and K produced the best results.

3.2 Real Data

3.2.1 Using Real World Data

The real world data used in this study was provided by the `CDM` package. This data came from the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA). The TIMSS exam is an assessment administered by the International Association for the Evaluation of Educational Achievement every four years, starting in the year 1995. This exam is administered to roughly 500,000 4th grade and 8th grade students in more than 50 countries to test math and science skills. In this study, a subset of the data from Austrian 4th graders from the 2007 version of the exam was used. The PISA exam is an assessment administered by the Organisation for Economic Co-operation and Development every three years, starting in the year 2000. This exam is administered to at least a few hundred thousand 15-year old students in 80 countries worldwide to test reading, math, and science skills. In this study, a subset of the data from German students from the 2000 version of this exam was used.

From the `CDM` package, a Q matrix and X matrix were provided for both the PISA 2000 and the TIMSS 2007 datasets. The X matrix for each additionally contained a column which indicated the subgroup membership for each examinee. In this case, the subgroup was gender, meaning participants belonged either to the “female” or “male” subgroup. This column was extracted and became the group assignment vector.

From here, `CDM::gdina` was used to construct a DINA model from this data. Then, `CDM::gdina.dif` was used to produce the DIF object. In the `GDINA` package, only `GDINA::dif` was used to produce the DIF object; the necessary matrices were provided, and the model type (DINA) was specified as an argument to this function. The type of statistical test, Wald or LR, was specified as well. Both the Wald and LR tests were used to flag items displaying DIF, in the same manner as for the simulated data. For this real-world data, if an item was identified with DIF, this indicated that the program has identified that the gender of the examinee played a part in their success on that test item, not just their possession of underlying latent attributes. In the case of real-world data, there was no set of true labels that could be used for comparison.

Since these real-world tests did not have a defined list of test items that display DIF, as is the case with the simulated data, further analysis was conducted on the examinees and test items to decide if the Wald and LR tests had performed well. The datasets included data beyond the test results, such as the attribute prevalence for the gender subgroups (addressed more in Sections 3.2.2 and 3.2.3). Finding trends in this additional data revealed if these tests have properly identified DIF. For example, one subgroup could have been more likely to possess certain attributes which are required to succeed on a handful of test items. The results of these tests should then have shown that this subgroup had a higher success rate on those items.

As described above, each real-world dataset provided the Q matrix, X matrix, attribute prevalence for each gender subgroup, and formal designation of each examinee into one gender subgroup. Due to the access to the X matrix and group membership, it was possible to break down the differing success rates of different subgroups on different items and compare this with the results of the Wald and LR methods, as seen in Sections 3.2.2 and 3.2.3. Due to the access to the Q matrix, it was also possible to see which attributes were necessary for success on each test item.

The `CDM` package implementation of the Wald test had more robust results than that of the `GDINA` package (see Section 3.1.4 for more elaboration), so the DIF object created from the `CDM::gdina.dif` function was used rather than that created from the `GDINA::dif` function for all real-world data assessments besides the Wald Test Results described in Sections 3.2.2 and 3.2.3.

3.2.2 TIMSS Analysis

Understanding the TIMSS Dataset This subset of data came from the TIMSS (Trends in Mathematics and Science Study) 2007 exam, a math exam administered to a large group of fourth grade students in Austria. The TIMSS 2007 dataset contains 698 examinees, 25 test items, and 15 underlying attributes. There are 333 female examinees and 365 male examinees according to the binary assignments provided by this dataset. This dataset's item response

(X) matrix, which provides the binary representation of which examinees were successful on which items, contains many "not available" ("NA") values. These were left untouched while running analyses. According to Lee et al., an "NA" entry in the X matrix corresponded to an "omitted or unreached" item, implying that any items that did not have a binary result were simply not answered by the student due to time or other factors [Lee et al., 2011].

Wald Test Results The dataset provided a pre-constructed Q matrix and item success (X) matrix. In this response matrix, one column was included to indicate the binary subgroup division of the examinees. This group membership vector was extracted. Just as with the simulated data, there are two methods by which the Wald test can be performed. First, the three aforementioned data structures were passed into the `CDM::gdina` function in order to create a DINA model. Then, as with the simulated data, a DIF object was created using `CDM::gdina.dif`.

An item-wise Wald test was performed and p-values were produced in order to indicate which items might display bias. Again, an item was considered significant if its associated p-value was equal to or less than the threshold $\alpha = 0.05$. Of the 25 test items in the TIMSS dataset, the Wald test provided by the `CDM` DIF-identifying function produced 1 significant p-value. This was item 12. The name of this item is M041275.

Using the same Q matrix, X matrix, and group membership vector, a DINA model and subsequent DIF object were created using the `GDINA` package. The `GDINA::dif` function was run twice, once with the Wald test setting and once with the LR test setting. P-values were examined in order to determine which items the statistical methods identified with bias.

Of the 25 test items in the TIMSS dataset, the Wald test implemented in `GDINA` produced 2 significant p-values. These were items 10 and 19. The names of these items are M041258B and M031242B. There was no overlap with the results of the Wald test from the `CDM` function.

Since the `CDM` implementation of the Wald test is preferred to the `GDINA`, the final result for the Wald test on the TIMSS data is that item M041275 was the only item flagged as displaying DIF.

Likelihood Ratio Test Results Of the 25 test items in the TIMSS dataset, the LR test executed by the `GDINA` DIF-identifying function produced 3 significant p-values. These were items 19, 20, and 21. The names of these items are M031242B, M031242C, and M031247. There was no overlap with the results of the Wald test via the `CDM` function, and one similar with the results of the Wald test via the `GDINA` package.

Benjamini-Hochberg Procedure The Benjamini-Hochberg (BH) Procedure was included to introduce the concept of controlling for false positives. False positives in the context of this research indicate test items that were incorrectly said to display DIF. The procedure was discussed in greater detail in Section 2.2.3.

The methodology of the BH Procedure was applied to each of the 25 items in the TIMSS data using the p-values obtained from the Wald test provided by the `CDM::gdina.dif` function. This test lowered, if necessary, the false positive rate of the original output to 10%.

Item	Wald p-val	Rank	BH critical value
M041275	0.0193	1	0.004
M041281	0.0898	2	0.008
M031247	0.1965	3	0.012
M041069	0.2140	4	0.016
M031303	0.2191	5	0.020
M041131	0.2429	6	0.024
M031242B	0.2950	7	0.028
M041164	0.3339	8	0.032
M031242A	0.4133	9	0.036
M041052	0.6401	10	0.040
M031245	0.6923	11	0.044
M031173	0.7231	12	0.048
M041146	0.7479	13	0.052
M041258A	0.7602	14	0.056
M031172	0.7757	15	0.060
M041186	0.7923	16	0.064
M031309	0.8070	17	0.068
M031085	0.8071	18	0.072
M031219	0.8092	19	0.076
M041258B	0.8548	20	0.080
M031242C	0.8920	21	0.084
M041336	0.9426	22	0.088
M041152	0.9692	23	0.092
M041076	0.9891	24	0.096
M041056	1.0000	25	0.1

Table 7: Benjamini-Hochberg Procedure, TIMSS

The final step of the BH Procedure was to find the largest p-value that was still smaller than the critical value. With this dataset, as seen in Table 7, there was no such item. This means that the BH Procedure had determined that no items display DIF when the examinees were grouped by gender and the false discovery rate was controlled at 10%. This additionally implies that the Wald test produced a result with a false discovery rate greater than 10%.

Mantel-Haenszel Test The Mantel-Haenszel (MH) Test, described further in Section 2.2.4, introduced a new type of test that can be run on this real-world data. Unlike the Wald and LR tests, the MH test relied on counts obtained from the data. 25 contingency matrices were constructed with simple counts of item success among the subgroups, one for each item. An example is shown in Table 8 below. These counts came from the X matrix obtained from the data. This matrix is an $N \times J$ matrix, as described in Section 2.1, where each cell indicates whether that examinee answered that test item correctly. This X matrix was divided into two new matrices, one for each gender subgroup, and the item success counts were obtained from these two matrices in order to populate these contingency tables.

Item M041052	**Correct**	**Incorrect**
Male	149	216
Female	118	215

Table 8: Subgroup success on Item M041052, TIMSS

The MH test takes K contingency tables, so in this case all 25 contingency tables were passed in. The χ^2 test statistic with 1 degree of freedom was calculated, and a p-value was given. If the p-value was statistically significant, this indicated that the gender subgroup did appear to have an effect on general item success.

Dataset	**χ^2 statistic**	**p-value**
TIMSS	2.3713	0.1236

Table 9: Mantel-Haenszel Test, TIMSS

The result displayed in Table 9 indicated that gender subgroup did not effect the outcome, as the p-value was not statistically significant. This MH test, however, did not give an indication of the specific items which may have been affected by gender. The provided statistic showed only that the test as a whole appeared to be unaffected by gender. This was where the McNemar test was introduced.

McNemar Test The McNemar test, described further in Section 2.2.5, takes one contingency table at a time. These tables were the same tables used in the MH test, as shown in Table 8. The test provided a χ^2 test statistic with 1 degree of freedom and p-value, just as the MH test, for that individual test item. This showed item-wise which exam questions might have been affected by the gender subgroup.

Item Name	χ^2 **statistic**	**p-value**	**Item Name**	χ^2 **statistic**	**p-value**
M041052	28.171	1.111e-07	M041336	177.97	2.2e-16
M041056	211.04	2.2e-16	M031303	123.75	2.2e-16
M041069	313.48	2.2e-16	M031309	28.986	7.292e-08
M041076	206.52	2.2e-16	M031245	294.39	2.2e-16
M041281	33.674	6.518e-09	M031242A	17.664	2.636e-05
M041164	52.503	4.297e-13	M031242B	0	1
M041146	61.669	4.062e-15	M031242C	2.1441	0.1431
M041152	91.785	2.2e-16	M031247	335.78	2.2e-16
M041258A	95.544	2.2e-16	M031219	1.1527	0.283
M041258B	157.83	2.2e-16	M031173	37.897	7.457e-10
M041131	171.74	2.2e-16	M031085	4	0.0455
M041275	111.02	2.2e-16	M031172	19.266	1.137e-05
M041186	53.603	2.454e-13			

Table 10: McNemar Test for each item, TIMSS.

Interestingly, the item-wise McNemar test provided a significant p-value for almost all of the items (22 out of 25). This varied from the results of the MH test, which indicated that gender subgroup was not significant for the test as a whole.

Attribute prevalence in subgroups The DIF object constructed from the DINA model in the `CDM` package allowed for further insight into the examinees. Specifically, the prevalence of each latent attribute was available per subgroup. Using this information, it was possible to examine how prevalent each underlying attribute was among the gender subgroups. If one attribute was held significantly more by one subgroup over another, it was possible that that group would succeed more on the items that require this attribute.

Attribute	Male	Female
1: Representing, comparing, and ordering whole numbers as well as demonstrating knowledge of place value.	0.490	0.463
2: Recognize multiples, computing with whole numbers using the four operations, and estimating computations.	0.936	0.844
3: Solve problems, including those set in real life contexts (for example, measurement and money problems).	0.835	0.845
4: Solve problems involving proportions.	0.575	0.412
5: Recognize, represent, and understand fractions and decimals as parts of a whole and their equivalents.	0.311	0.309
6: Solve problems involving simple fractions and decimals including their addition and subtraction.	0.347	0.322
7: Find the missing number or operation and model simple situations involving unknowns in number sentence or expressions.	0.314	0.471
8: Describe relationships in patterns and their extensions; generate pairs of whole numbers by a given rule and identify a rule for every relationship given pairs of whole numbers.	0.756	0.686
9: Measure, estimate, and understand properties of lines and angles and be able to draw them.	0.538	0.566
10: Classify, compare, and recognize geometric figures and shapes and their relationships and elementary properties.	0.452	0.453
11: Calculate and estimate perimeters, area, and volume.	0.489	0.575
12: Locate points in an informal coordinate to recognize and draw figures and their movement.	0.686	0.580
13: Read data from tables, pictographs, bar graphs, and pie charts.	0.687	0.622
14: Comparing and understanding how to use information from data.	0.745	0.726
15: Understanding different representations and organizing data using tables, pictographs, and bar graphs.	0.585	0.593

Table 11: Attribute prevalence for each gender, TIMSS

These attributes, as shown in Table 11, appeared relatively consistent between gender subgroups. Attributes 2, 4, 8, and 12 had a notably higher prevalence among the male subgroup, while attributes 7 and 11 had a higher prevalence among the female subgroup.

Attributes in test items The Q matrix, a $J \times K$ matrix, provided the information of which underlying latent attributes were necessary to succeed on each item. This information was consolidated in the table below.

No.	Item Name	Attributes	No.	Item Name	Attributes
1	M041052	1,2	14	M041336	1,2,5,6,13,14
2	M041056	5	15	M031303	2,3
3	M041069	2,4,5	16	M031309	2,3
4	M041076	3,6	17	M031245	2,7
5	M041281	2,3,8	18	M031242A	2,3,8
6	M041164	10,12	19	M031242B	2,3,14
7	M041146	9,10,12	20	M031242C	2,3,8,14
8	M041152	1,2,3,10,11	21	M031247	2,3,7
9	M041258A	10	22	M031219	10,11,12
10	M041258B	9,10	23	M031173	2,3
11	M041131	2,3,4,9	24	M031085	10
12	M041275	1,13,15	25	M031172	1,2,13,15
13	M041186	1,2,4,13			

Table 12: Attributes necessary for each test item, TIMSS.

Taking into consideration the prevalence of the attributes, it might be expected that the male subgroup had greater success on items requiring attributes 2, 4, 8, and 12 while the female subgroup had greater success on items requiring attributes 7 and 11. Males may have been better at items 3 (M041069), 5 (M041281), 11 (M041131), 18 (M031242A), 19 (M031242B), and 20 (M031242C); females may have been better at items 17 (M031245) and 22 (M031219).

Group success on test items From looking at the data matrix from the real world datasets, the difference in performance between the subgroups could be identified. Items with significant differences in performance may display bias. These items could be cross checked with the system-identified items in order to see if the model had picked up on the same items that were selected manually. In the table below, the percentages of each subgroup that found success on each item are indicated. Any items with a 5% difference or higher in performance are highlighted in yellow.

Item	Female	Male	Item	Female	Male
M041052	35.44%	40.82%	M041336	15.32%	16.99%
M041056	11.41%	15.07%	M031303	78.08%	83.56%
M041069	3.30%	4.93%	M031309	66.97%	66.58%
M041076	12.31%	14.52%	M031245	5.11%	5.75%
M041281	34.83%	38.63%	M031242A	60.66%	65.75%
M041164	34.23%	30.41%	M031242B	53.15%	51.78%
M041146	31.53%	30.14%	M031242C	55.26%	57.26%
M041152	25.83%	26.85%	M031247	2.40%	1.10%
M041258A	26.73%	24.11%	M031219	55.26%	55.34%
M041258B	18.92%	16.44%	M031173	66.67%	70.14%
M041131	14.71%	20.00%	M031085	48.35%	45.21%
M041275	21.62%	26.85%	M031172	64.86%	63.56%
M041186	29.43%	36.44%			

Table 13: Subgroup success on each item, TIMSS.

Six items had a difference in success of 5% or greater between gender subgroups. All of these items showed the male subgroup performing higher than the female subgroup. These items required the attributes 1, 2, 3, 4, 8, 9, 13, and 15. All of these attributes except attributes 3 and 9 had higher prevalence in the male subgroup, therefore it was not unusual that the male subgroup performed higher on these test items.

To refer back to the items marked as potentially biased in the previous section, the male subgroup did indeed perform better than the female subgroups on item 11 (M041131) and item 18 (M031242A).

For comparison, the Wald test from the `CDM` package identified one of these items (item M041275); the Wald test and the LR test from the `GDINA` package did not identify any of these items. In this case, the Wald test provided by the `CDM` package performed better than LR test at flagging these manually identified items, though not by much.

3.2.3 PISA Analysis

Understanding the PISA Dataset This subset of the PISA (Programme for International Student Assessment) 2000 exam came from a reading test administered to a large group of 15-year-old German students. This test examines reading, math, and science knowledge. The PISA 2000 dataset contains 1095 examinees, 26 test items, and 6 underlying attributes. Of the 1095 examinees, 525 are labeled as male and 570 as female.

Wald Test Results Just as with the TIMSS dataset, the PISA dataset came with a Q matrix and an X matrix which also contained gender subgroup membership for each examinee. These were used to construct a DINA model using `CDM::gdina`, which was then used to construct a DIF object. Of the 26 test items in the PISA dataset, the Wald test from the `CDM` DIF-identifying function produced 8 significant p-values ($\leq \alpha = 0.05$). These were items 1, 2, 7, 10, 11, 16, 17, and 24. The names of these items are R040Q02, R040Q03A, R077Q03, R077Q06, R088Q01, R110Q01, R110Q04, and R216Q06.

A DINA model was constructed, followed by a DIF object. Of the 26 test items in the PISA dataset, the Wald test implemented by the `GDINA` DIF-identifying function produced 14 significant p-values. These were items 1, 2, 6, 7, 9, 10, 16, 17, 18, 20, 21, 22, 24, and 25. The names of these items are R040Q02, R040Q03A, R077Q02, R077Q03, R077Q05, R077Q06, R110Q01, R110Q04, R110Q05, R216Q01, R216Q02, R216Q03T, R216Q06, and R236Q01. 7 of these overlapped with the results of `CDM` Wald test.

Likelihood Ratio Test Results Of the 26 test items in the PISA dataset, the `GDINA` DIF-identifying function produced 15 significant p-values, using the LR test. These were items 1, 2, 6, 7, 9, 10, 16, 17, 18, 20, 21, 22, 24, 25, and 26. The names of these items are R040Q02, R040Q03A, R077Q02, R077Q03, R077Q05, R077Q06, R110Q01, R110Q04, R110Q05, R216Q01, R216Q02, R216Q03T, R216Q06, R236Q01, and R236Q02. 7 of these overlapped with the results of the Wald test from the `CDM` function, and 14 of these overlapped with the results of the Wald test from the `GDINA` package.

Benjamini-Hochberg Procedure The Benjamini-Hochberg (BH) Procedure introduced the concept of controlling for false positives. False positives in the context of this research indicate test items that were incorrectly said to display DIF. The procedure is discussed in greater detail in Section 2.2.3.

The methodology of the BH Procedure was applied to each of the 26 items in the PISA data using the p-values obtained from the Wald test provided by the `CDM::gdina.dif` function.

Item	Wald p-val	Rank	BH critical value
R040Q03A	0.0000	1	0.003846
R216Q06	0.0000	2	0.007692
R040Q02	0.0001	3	0.01154
R077Q03	0.0002	4	0.01538
R110Q04	0.0003	5	0.01923
R110Q01	0.0055	6	0.02308
R088Q01	0.0449	7	0.02692
R077Q06	0.0486	8	0.03077
R110Q05	0.0978	9	0.03462
R216Q01	0.1447	10	0.03846
R077Q05	0.1706	11	0.04231
R236Q02	0.2525	12	0.04615
R077Q04	0.2641	13	0.05
R040Q03B	0.2915	14	0.05385
R216Q04	0.3079	15	0.05769
R088Q04T	0.3539	16	0.06154
R236Q01	0.3934	17	0.06538
R088Q07	0.3997	18	0.06923
R040Q04	0.4805	19	0.07308
R040Q06	0.5318	20	0.07692
R077Q02	0.6116	21	0.08077
R216Q02	0.6771	22	0.08462
R216Q03T	0.6814	23	0.08846
R088Q05T	0.7753	24	0.09231
R088Q03	0.8731	25	0.09615
R110Q06	0.9794	26	0.1

Table 14: Benjamini-Hochberg Procedure, PISA

The largest p-value that was still smaller than the critical value was the item of Rank 6. This means that the BH Procedure determined that the first 1-6 items listed here (all those above the dashed line in Table 14) were statistically significant and displayed DIF when the examinees were grouped by gender and the false discovery rate was controlled at 10%. Comparatively, the Wald test originally identified the first 8 items to display DIF. This implies that the Wald test produced a result with a false discovery rate greater than 10%.

Mantel-Haenszel Test 26 contingency matrices are constructed with simple counts of item success among the subgroups, one for each item. An example is shown in Table 15 below.

Item R040Q02	**Correct**	**Incorrect**
Male	340	185
Female	327	243

Table 15: Subgroup success on Item R040Q02, PISA

The MH test took K contingency tables, so in this case all 26 contingency tables were passed in. The χ^2 test statistic with 1 degree of freedom was calculated, and a p-value was given. If the p-value was statistically significant, this indicated that the gender subgroup did appear to have an effect on item success in the exam as a whole.

Dataset	**χ^2 statistic**	**p-value**
PISA	174.58	2.2e-16

Table 16: Mantel-Haenszel Test, PISA

The results in Table 16 showed that gender did appear to have a significant effect on the outcome of the test as a whole.

McNemar Test The McNemar test took one contingency table at a time, and provided a χ^2 test statistic with 1 degree of freedom and p-value, just as the MH test, for that individual test item. This showed item-wise which exam questions might have been affected by gender subgroup.

Item Name	χ^2 statistic	p-value	Item Name	χ^2 statistic	p-value
R040Q02	38.83	4.623e-10	R088Q05T	49.295	2.202e-12
R040Q03A	3.1168	0.07749	R088Q07	11.391	0.0007382
R040Q03B	55.63	8.747e-14	R110Q01	198.07	2.2e-16
R040Q04	143.51	2.2e-16	R110Q04	182.05	2.2e-16
R040Q06	12.115	0.0005001	R110Q05	60.137	8.848e-15
R077Q02	124.38	2.2e-16	R110Q06	109.6	2.2e-16
R077Q03	14.565	0.0001354	R216Q01	95.208	2.2e-16
R077Q04	8.2086	0.004169	R216Q02	2.8045	0.094
R077Q05	98.799	2.2e-16	R216Q03T	18.348	1.84e-05
R077Q06	5.2354	0.02213	R216Q04	35.136	3.075e-09
R088Q01	35.705	2.296e-09	R216Q06	45.486	1.537e-11
R088Q03	47.108	6.718e-12	R236Q01	1.55	0.2131
R088Q04T	19.406	1.057e-05	R236Q02	161.16	2.2e-16

Table 17: McNemar Test for each item, PISA.

Similar to the TIMSS data, a majority of the items were found to be statistically significant (23 of the 26 items). This aligns with the result of the MH test, which shows that gender did have an effect on the test outcomes.

Attribute prevalence in subgroups To get a better sense of how the attributes presented in each gender group, attribute prevalence was examined. Each of the 6 attributes are listed below, followed by the prevalence in each of the 2 genders in this study. This was found by looking at the skill pattern and group statistics provided by the gender-based `CDM` DINA model generated earlier. From Table 18, every attribute except 4 had higher prevalence in the female subgroup.

Attribute	Male	Female
1: Locating Information	0.636	0.752
2: Forming a broad, general, understanding	0.696	0.825
3: Developing a logical interpretation	0.646	0.710
4: Evaluating a number-rich text with number sense	0.621	0.611
5: Evaluating the quality or appropriateness of a text	0.631	0.733
6: Test speededness	0.492	0.757

Table 18: Attribute prevalence for each gender, PISA.

The next step was to break down which items on the PISA test required certain attributes. If success on one item was based heavily on an attribute that the female demographic group was much more likely to possess, it might be expected to see this test item displayed DIF.

Attributes in test items Next, it was relevant to see which attributes were deemed necessary for success on each test item. This information was extracted from the Q matrix provided with the data. After discovering which attributes were needed for particular test items, certain items could be flagged as potentially biased. For example, if a test item required 3 attributes most commonly held by the male subgroup, this item may have favored the success of the male subgroup. These potentially biased items could then be compared with the output of the statistical tests.

No.	**Item Name**	**Attributes**	**No.**	**Item Name**	**Attributes**
1	R040Q02	3,4	14	R088Q05T	2,4
2	R040Q03A	1,4	15	R088Q07	2,4,5
3	R040Q03B	4,5	16	R110Q01	2,5
4	R040Q04	2,4	17	R110Q04	1,3
5	R040Q06	3,4	18	R110Q05	1,3
6	R077Q02	1	19	R110Q06	1,3
7	R077Q03	2,5	20	R216Q01	2,6
8	R077Q04	3,5	21	R216Q02	3,5,6
9	R077Q05	2,5	22	R216Q03T	1,3,6
10	R077Q06	1,3	23	R216Q04	3,6
11	R088Q01	2,4	24	R216Q06	1,3,6
12	R088Q03	1,4	25	R236Q01	1,3,6
13	R088Q04T	3,4	26	R236Q02	3,6

Table 19: Attributes necessary for each test item, PISA. Highlight indicates difference of at least 5%.

Taking into consideration the prevalence of the attributes, it might be expected that the female subgroup had greater success on items requiring attributes 1, 2, 5, and 6. Females may have been better at items such as 7 (R077Q03), 9 (R077Q05), 20 (R216Q01), and 21 (R216Q02).

Group success on test items In the table below, the percentages of each subgroup that found success on each item are indicated. Just as with the TIMSS data in the previous section, any items with a 5% difference or higher in performance are highlighted in yellow. Any items with a 10% difference or higher are highlighted in green. Finally, any items with a 20% difference or higher are highlighted in pink.

Item	Female	Male	Item	Female	Male
R040Q02	57.37%	64.76%	R088Q05T	63.68%	62.67%
R040Q03A	46.32%	57.33%	R088Q07	57.72%	52.95%
R040Q03B	31.93%	32.19%	R110Q01	85.44%	74.29%
R040Q04	76.49%	72.19%	R110Q04	85.61%	72.00%
R040Q06	55.96%	55.05%	R110Q05	70.70%	60.00%
R077Q02	77.54%	68.38%	R110Q06	73.51%	68.57%
R077Q03	65.26%	47.81%	R216Q01	80.70%	60.57%
R077Q04	57.02%	51.43%	R216Q02	58.95%	44.19%
R077Q05	30.00%	21.52%	R216Q03T	45.26%	30.29%
R077Q06]	47.54%	37.52%	R216Q04	38.42%	30.86%
R088Q01	65.26%	57.14%	R216Q06	75.26%	52.00%
R088Q03	63.86%	61.90%	R236Q01	51.58%	37.90%
R088Q04T	60.18%	55.05%	R236Q02	24.05%	15.43%

Table 20: Subgroup success on each item, PISA. Yellow highlight indicates difference between 5 and 10%, green indicates between 10 and 20%, pink indicates greater than 20%

19 out of the 26 PISA items, a majority, had at least a 5% difference in performance between the gender subgroups. The item on which the male subgroup performed higher is R040Q02; the female subgroup outperformed the male subgroup on the remainder of the highlighted items. The items with a performance difference between 5% and 10% required different combinations of all the attributes. The items with a performance difference between 10% and 20% also required different combinations of all the attributes. Finally, the items with a performance difference of more than 20% required the attributes 1, 2, 3, and 6. All attributes except attribute 4 were found in higher prevalence in the female subgroup, so it is not unusual to see that the female subgroup outperformed the male group on almost all highlighted items. In particular, much more of the female subgroup possessed attribute 6, explaining the larger discrepancy in performance for the pink-highlighted items. Referring back to the items flagged as potentially bias in the previous section, females did in fact perform better on item 7 (R077Q03), 9 (R077Q05), 20 (R216Q01), and 21 (R216Q02).

For comparison, the Wald test conducted via the `CDM` package identified eight of these items (items R040Q02, R040Q03A, R077Q03, R077Q06, R088Q01, R110Q01, R110Q04, and R216Q06). The Wald test conducted via the `GDINA` package identified 14 of these items (R040Q02, R040Q03A, R077Q02, R077Q03, R077Q05, R077Q06, R110Q01, R110Q04, R110Q05, R216Q01, R216Q02, R216Q03T, R216Q06, and R236Q01). The LR test conducted via the GDINA package identified 15 of these items (R040Q02, R040Q03A, R077Q02, R077Q03, R077Q05, R077Q06, R110Q01, R110Q04, R110Q05, R216Q01, R216Q02, R216Q03T, R216Q06, R236Q01, and R236Q02). In this case, the results of the LR test aligned the most with the manually identified biased items, closely followed by the Wald test provided by the GDINA package.

4 Discussion

After an analysis of the Wald and LR tests on simulated data, the Wald test, specifically as implemented by the CDM package, appeared to have more robust results. With the simulated data, it was known precisely which items were manipulated to be biased and therefore it was simple to see how well the Wald and LR tests performed. This was a supervised technique, meaning the true labels (DIF or no DIF) were known. In this scenario, a false negative was the most egregious error for the model to make, as this means biased items could go unidentified and continue to perpetuate unfair testing for certain demographic groups. To compare the performance of the Wald and LR tests, it was important to compare the false negative and false positive rates. False positives were additionally important because if the model over-identified DIF, i.e. signaled that every item was biased, the results became meaningless. The Wald test outperformed the LR test in the simulated data, as seen in Section 3.1.4. Notably, the LR test had a very high FPR, or overidentification of DIF, when the sample size was large; this may be in part due to increased sensitivity as the sample size increases. Of the Wald implementations, the CDM package performed better than the GDINA package. This is likely because the CDM packages had more control over parameter estimation. At the highest setting, there were 2^9 parameters to estimate (2 subgroups, 9 underlying attributes), which was very large.

Although the "true" answers were unknown, examining the data provided by the CDM package gave a well-informed view on which test items may have displayed bias. Many more items were identified as biased in the PISA 2000 dataset than the TIMSS 2007 dataset. Simply due to the nature of the exams, it made sense that there might be more bias in the PISA 2000 exam than the TIMSS 2007 exam, as PISA involves reading and TIMSS is more mathematical. Additionally, the age group these exams test may have been significant. 4th graders (TIMSS) are much younger than 15-year-olds (PISA), and as students age and progress, questions become more complicated; this means older groups are more susceptible to bias than younger children.

To make some recommendations for practitioners, for any of the described DIF-identification methods, a larger sample size provides better results. More specifically, with a sufficiently large sample size the Wald test yields the most favorable results, as has been seen in other studies such as [Hou et al., 2014]. If a user is concerned of the possibility of over-identification, the BH procedure will aid in controlling the FPR. If examining a test with a large amount of latent attributes, of the methods examined in this study, the Wald test is likely to give the best performance. However, more complex bias-identification is recommended. This also applies for practitioners examining more than 40 items at a time. Finally, it is not recommended to us the MH and McNemar tests as steadfast indicators of bias when the test item is indicated as the strata. These tests are simple baselines, and while they can provide a starting point for identifying DIF in test items, they do not provide a nuanced enough examination of test items as they rely only on the success counts of each subgroup and tended toward over-identification in this study.

There are limitations to this study, including the choice to examine a binary subgroup. A multi-class subgroup such as race or country of origin would introduce another layer of complexity to this study. This would be a future direction of reviewing DIF-identification methods. Additionally, in this study the MH test is performed with test item as the third grouping variable; another angle would be to use the latent attribute profile as the strata. Introducing more complex models, such as neural networks and other machine learning techniques, may also allow for models to finetune DIF-identification and alleviate some of the limitations of simpler statistical methods. Additionally, the proportion of DIF items was held constant at one fifth in each simulation, and it would be important to examine if varying the proportion of biased items affects the performance of these methods. Finally, introducing other types of CDMs in addition to DINA would be another further direction this research could take.

As a whole, these statistical methods provide good starting points for the problem of DIF-identification. There are certain limitations to the use of each test described in this study. Therefore, this important discussion of equitizing standardized testing across the globe may benefit from learning more complex models.

References

[de la Torre, 2011] de la Torre, J. (2011). The generalized dina model framework. *Psychometrika*, 76(2):179–199.

[Hou et al., 2014] Hou, L., de la Torre, J., and Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the wald test to investigate dif in the dina model. *Journal of Educational Measurement*, 51(1):98–125.

[Lee et al., 2011] Lee, Y.-S., Park, Y. S., and Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in massachusetts, minnesota, and the u.s. national sample using the timss 2007. *International Journal of Testing*, 11(2):144–177.

[Ma and de la Torre, 2020] Ma, W. and de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, 93(14):1–26.

[Ma et al., 2021] Ma, W., Terzi, R., and de la Torre, J. (2021). Detecting differential item functioning using multiple-group cognitive diagnosis models. *Applied Psychological Measurement*, 45(1):37–53.

[Mehrazmay et al., 2021] Mehrazmay, R., Ghonsooly, B., and de la Torre, J. (2021). Detecting differential item functioning using cognitive diagnosis models: Applications of the wald test and likelihood ratio test in a university entrance examination. *Applied Measurement in Education*, 34(4):262–284.

[R Core Team, 2021] R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Robitzsch et al., 2020] Robitzsch, A., Kiefer, T., George, A. C., and Ünlü, A. (2020). *CDM: Cognitive Diagnosis Modeling*. R package version 7.5-15.

[Wainer and Sireci, 2005] Wainer, H. and Sireci, S. G. (2005). Item and test bias. *Encyclopedia of Social Measurement*, pages 365–371.