***Dirasat:*** *Educational Sciences*, Volume 51, No. 3, 2024

Deanship of Scientific Research

# The Effectiveness of Mantel Haenszel Log Odds Ratio Method in Detecting Differential Item Functioning Across Different Sample Sizes and Test Lengths Using Real Data Analysis

***Reem Mohammad Elyan*[1], *Majed Mahmoud Al jodeh*[2]***

[1]Deanship of Scientific Research and Postgraduate Studies, Yarmouk University, Irbid, Jordan

[2]Department of Education and Psychology, College of Education and Arts, University of Tabuk, Tabuk, Saudi Arabia

Received: 7/2/2024
Revised: 31/3/2024
Accepted: 29/5/2024
Published: 15/9/2024

* *Corresponding author:*
majed_jodeh@hotmail.com

Citation: Elyan, R. M. ., & Al jodeh, M. M. . (2024). The Effectiveness of Mantel Haenszel Log Odds Ratio Method in Detecting Differential Item Functioning Across Different Sample Sizes and Test Lengths Using Real Data Analysis. *Dirasat: Educational Sciences*, *51*(3), 37–46. https://doi.org/10.35516/edu.v51i3.6755



**Abstract**

**Objectives:** This study aims to determine the effectiveness of the Mantel Haenszel Log Odds Ratio method in detecting Differential Item Functioning (DIF) across gender, while considering variations in sample size and test length. Utilizing real data, the study draws from a sample of tenth-grade students in Jordan who participated in the 2018 PISA International Mathematics Test.

**Methods:** The study employs the experimental methodology, utilizing three levels of sample size and test length: (342, 200, and 100) and (30, 20, and 10), respectively. Nine iterations of the DDFS program were conducted to collect the results, representing nine scenarios resulting from the intersection of sample size and test length levels.

**Results:** The study indicates that variations in sample size and test length significantly affect the Mantel-Hanzel (MH) method. Specifically, it observes an improvement in the MH method's ability to detect DIF items with larger sample sizes, while maintaining a consistent test length. Conversely, the method's efficacy declines with longer test lengths, despite maintaining a fixed sample size at a specific level.

**Conclusion:** The study recommends using a large sample size and a short test length for effective detection of DIF items using the MH method.

**Keywords**: Mantel Haenszel, Log Odds Ratio, DIF, Real Data, PISA test, Tenth-grade.

# فاعلية طريقة نسبة الأرجحية لمانتل هانزل لاكتشاف الأداء التفاضلي للفقرة باختلاف حجم العينة وطول الاختبار باستخدام بيانات حقيقة

***ريم محمد عليان*[1] ، *ماجد محمود الجوده*[2]***

[1] عمادة البحث العلمي والدراسات العليا، جامعة اليرموك، إربد، الأردن

[2] قسم التربية وعلم النفس، كلية التربية والآداب، جامعة تبوك، تبوك، المملكة العربية السعودية

**ملخّص**

**الأهداف:** هدفت الدراسة الحالية لدراسة أداء طريقة نسبة الأرجحية لمانتل هانزل لاكتشاف الأداء التفاضلي للفقرة تبعاً لمتغير الجنس، باختلاف حجم العينة، وطول الاختبار باستخدام بيانات حقيقية تم الحصول عليها من استجابات طلبة الصف العاشر في الأردن على اختبار بيزا الدولي 2018 في مبحث الرياضيات.

**المنهجية:** تم استخدام المنهج التجريبي في هذه الدراسة، حيث تم استخدام ثلاثة مستويات لحجم العينة، وطول الاختبار (342، 200، 100)، (30،20،10)، وتم تنفيذ برنامج DDFS تسع مرات، لإنجاز نتائج تسع حالات ناتجة من تقاطع مستويات حجم العينة، وطول الاختبار.

**النتائج:** توصلت الدراسة إلى أن التغيير في حجم العينة وطول الاختبار يؤثر بشكل كبير على هذه الطريقة، وأن فعاليتها في الكشف عن الأداء التفاضلي للفقرات تتحسن بزيادة حجم العينة K وبقاء طول الاختبار ثابتاً عند مستوى معين، وبتضائل مع زيادة طول الاختبار على الرغم من الحفاظ على حجم عينة ثابت عند مستوى معين.

**الخلاصة:** إن النتيجة التي توصلت إليها هذه الدراسة، أنه إذا تم استخدام هذه الطريقة لاكتشاف الأداء التفاضلي للفقرات، توصي باستخدام حجم عينة أكبر، وطول اختبار قصير، لأن فعاليتها تزداد في هذه الحالة.

**الكلمات الدالة :**مانتل هانزل، نسبة الأرجحية، الأداء التفاضلي، بيانات حقيقية، اختبار بيزا، الصف العاشر.

## 1. INTRODUCTION

Measures and tests are necessary to achieve a number of goals, such as individual classification, university admissions, and evaluating characteristics like IQ and attitudes. Fairness and objectivity are the most critical factors in test choices, which require regular item review to ensure test validity of diverse subgroups of examinees. As Differential Item Functioning (DIF) can negatively affect test conclusions and classification of individuals. Several studies investigated it on achievement tests such as the TOEFL (Test of English and the Scholastic Assessment Test), MELAB (Michigan English Language Assessment Battery) (Park, 2008; Wagner,2004; Eom, 2008; Vahid et al., 2011). National and international tests are essential for evaluating educational systems worldwide. These tests measure student proficiency in basic subjects and abilities, including reading, arithmetic, and science, essential for improving education (Gu, 2023; Ihlenfeldt & Rios, 2023; Wall & Horák, 2006, 2008).

One of the prominent tests on a global scale is the Program for International Student Assessment (PISA) examination affiliated with the Organization for Economic Cooperation and Development (OECD). Every three years, a study covering fifteen-year-old children in grades seven through twelve is conducted to see how well pupils are acquiring the knowledge and abilities needed to engage with society in all domains, including science, math, reading, and knowledge. This triennial exam compares student academic performance to a standard baseline. Regardless of curriculum, its main focus is on 15-year-old students and it evaluates their reading, science, and math skills. Because of this, some activities may help some student groups while having a negative impact on other student groups (Marôco, 2021; Münch & Wieczorek, 2023).

Since a few aspects may affect the DIF, It is crucial to investigate how DIF items perform in different contexts including how the sample size and test length relate to the gender variable.

### *1.1. Differential Item Functioning*

It is "a statistically derived function that expresses the difference in item response between two groups at the same ability level," according to Camilli & Shepard (1994). Williams (1997) distinguished between DIF and bias, showing that a test item is biased if it measures attributes unrelated to the test. A DIF occurs when subjects of equal ability from different subgroups have unequal probability values to respond correctly.

In test construction, the use of differential item functioning and the requirement to provide indicators of the measures' validity and reliability have grown in importance. The most recent version of the Standards for Educational and Psychological Testing recognizes Differential Item Functioning (DIF) as internal structure-based validity evidence.

The researchers distinguished between two types of DIF: uniform DIF (UNDIF) and nonuniform DIF (NUNDIF), emphasizing the relationship between group membership and ability level. According to Ackerman (1992), Item Characteristic Curves (ICCs) are parallel for UNDIF and nonparallel for NUNDIF (Ackerman, 1992; Mellenbergh, 1982, 1989; Millsap & Everson, 1993; Narayanon & Swaminathan, 1996). Some methods for detecting differential item functioning rely on some indicators to determine its type. For example, the Mantel Haenszel Log Odds Ratio Method used in this study relies on the BD indicator (Breslow-Day Chi-Square test): This indicator is used to demonstrate the presence of non-uniform DIF for items at the level of statistical significance of 0.05, and its critical values were 3.84 (Penfield, 2010).

### *1.2. Detecting Differential Item Functioning*

Numerous techniques, including area approaches, logistic regression (LR), Rasch model, multiple groups in Item Response Theory (IRT) model, Simultaneous Item Bias Tests (SIBTETSs), and Multiple Indicators Multiple Causes (MIMIC) models, have been developed to detect DIF. The Mantel-Haenszel Log-Odds-Ratio (MH-LOR) method is one of the most often used. It is predicated on a DIF investigation between two groups: the focal group (FG) and the reference group (RG) (Holland & Thayer, 1986).

The method concisely uses the Z test with a harmonic table (2 x 2) to compare correct and incorrect answers for items between FG and RG at each level of ability. The differences between the two groups are compared for each test item to determine the respective ability levels. The DIF is then determined, whether or not it is in favour of any group, by looking

at the value of the odds ratio($\hat{\alpha}_{MH}$) (Mantel & Haenszel, 1959). **Table 1** shows the distribution of the responses of the reference and focal group members for the item at the k level of ability.

**Table 1*:* Distribution of the responses of the members of reference and focal groups for the item at the k level of ability**

| Group | Correct answer (1) | False answer (0) | Total |
|---|---|---|---|
| **Reference** | $A_k$ | $B_k$ | $A_k + B_k$ |
| **Focal** | $C_k$ | $D_k$ | $D_k + C_k$ |
| **Total** | $A_k + C_k$ | $B_k + D_k$ | $T_k$ |

Where:

$A_k$: Number of respondents in the reference group who answered the item correctly.

$B_k$: Number of respondents in the reference group who did not answer the item correctly.

$C_k$: Number of respondents in the focal group who answered the item correctly.

$D_k$: Number of respondents in the focal group who did not answer the item correctly

**Equation 1:** The value of the odds ratio estimator is calculated to indicate DIF

$$\hat{\alpha}_{MH} = \frac{\sum A_k D_k / T_k}{\sum B_k C_k / T_k} \dots\dots (1)$$

If $\hat{\alpha}_{MH}$ =1, that means the item doesn't have DIF; if $\hat{\alpha}_{MH}$ > 1, then the item has DIF in favour of RF, and if $\hat{\alpha}_{MH}$ < 1, then the item has DIF in favour of the FG group (Penfield, 2010).

Based on **Table 1,** the Z test is used to test the null hypothesis:

$$H_0: \hat{\alpha}_{MH} = 1$$

By applying Equation 2*:*

$$Z = \frac{\sum [A_k - E(A_k)] - 0.5}{\sqrt{var(A_k)}} \dots\dots (2)$$

Where:

$E(A_k)$: The expected value of a number of subjects in RG who answered the item correctly.

$var(A_k)$: variance of $A_k$

***The values of each of them are calculated by applying the following*** *Equation 3 and Equation 4***:**

$$E(A_k) = \frac{(A_k + B_k)(A_k + C_k)}{A_k + B_k + C_k + D_k} \dots\dots (3)$$

$$var(A_k) = \frac{(A_k + B_k)(C_k + D_k)(A_k + C_k)(B_k + D_k)}{(A_k + B_k + C_k + D_k)^2 (A_k + B_k + C_k + D_k - 1)} \dots\dots (4)$$

To quantify the DIF of the item, in case the value of Z is statistically significant, a logarithmic transformation of the value of $\hat{\alpha}_{MH}$ is performed by the following **Equation 5**

$$\lambda_{MH} = ln\ ln\ (\hat{\alpha}_{MH}) \dots\dots\dots\dots\dots\dots\dots (5)$$

$\lambda_{MH}$: Known as Lambda Mantel Hansel

Based on the following criteria, a decision is made on the size of the DIF:

$|\lambda_{MH}| < 0.43$: Low DIF item $0.43 \leq |\lambda_{MH}| < 0.64$: Medium DIF item

$|\lambda_{MH}| \geq 0.64$: High DIF item (Penfield & Camilli, 2006)

### *1.3. Literature Review*

Based on an analysis of the theoretical literature on the subject of study into the impact of different variables on the MH method for determining DIF, the following studies are examined in this regard:

In a simulation research, sample size, test length, and DIF type were the three different conditions under which Swaminathan and Rogers (1990) compared the Mantel-Hansel approach with the LR method for determining the DIF (Uniform or Non-uniform) of the item. The findings show that when it comes to detecting DIF, the LR approach is more sensitive to test length and sample size than the MH method. More specifically, when utilizing LR, the percentage of DIF items rises with higher sample sizes and longer tests. Nonetheless, it is important to remember that the study's findings (Swaminathan & Rogers, 1990) indicate that the MH technique has limits when it comes to identifying non-uniform DIF.

Results of Finch's study (2005) indicated that the effectiveness of the DIF detection method increases at high-test lengths, specifically in the length of the 50-item test (Finch, 2005). In comparison to other approaches, the MH method's test power is the highest under conditions of sample size, test length, and DIF size, as demonstrated by the simulation study conducted by Kabasakala et al. (2014). Under those circumstances, this technique appears to have an impact on type I error. The most important aspect of this study is its use of correct data through a sample of PISA test results; other studies employed simulation data under pre-defined settings (Kabasakal et al., 2014).

Arikan et al. (2016) conducted a study investigating the similarities and differences in four methods for detecting DIF: MIMIC, SIBTEST, MH, and LR. MH in different sample sizes: 2000, 1200, 1000, 600, 300. The study concluded that the results are more effective in a sample size of 2000 or more, and the four methods are more consistent in detecting DIF items (Akın Arıkan et al., 2016).

Alomari et al. (2023) investigated the effect of sample size on the number of items with DIF using the MH method. Data was obtained from an 8th-class math exam with 40,000 strengths and 40 multiple-choice questions. At random, eight samples containing 250, 500, 1250, 2500, 5000, 10000, 15000, and 20000 respondents were chosen. This investigation showed that more items were found using DIF and DDF as the sample size increased. To find items with nonuniform DIF and DIF of insignificant magnitude, more significant sample sizes are also required (Alomari et al., 2023).

Some studies have indicated that the effectiveness of methods for detecting differential item functioning increases as the sample size increases, in a study for Dorans & Holland (1992)described the Mantel-Haenszel method and compared it with item response theory methods in terms of similarities and differences in procedures, and several issues in applied DIF analyses were discussed including inclusion of the studied item in the matching variable, and refinement of the matching variable, they suggested whenever feasible, the largest possible sample size for both focal and reference groups should be used in DIF , the results of (Gao, 2019) study, which compared six methods for detecting differential functioning, including Mantel-Haenszel method, concluded that all methods work well when the sample size increases.

### *1.4. Study Objectives*

The significant role that tests generally play in achieving many purposes and the student's fateful decisions based on them, such as those concerned with promoting students, classifying them, branching them out, and accepting them in various university majors, etc., give the importance of international assessment tests in providing fundamental indicators about students' skills and knowledge and the different learning and teaching contexts, the level and quality of education in educational systems and institutions, their significant role in providing in-depth information about the factors that affect the results of students, the development of their attitudes and skills, and the arrangement and classification of countries based on the results of their evaluation.

Researchers in this sector have been particularly interested in the DIF issue, which has a negative impact on test features and findings. The identification of DIF of items is a test-related problem that can be impacted by a number of variables, such as sample size, test length, participant characteristics, question phrasing, etc. In order to determine whether the MH method can correctly identify the DIF of gender for the item in various sample sizes and test lengths, this study will use a sample of tenth-grade students who took the PISA 2018 mathematics test.

The widespread use of the Mantel-Haenszel method in detecting differential item functioning among researchers on the

one hand, and the importance of studying differential item functioning on the other hand, increases the importance of this study, which came to examine the effectiveness of this method under conditions of varying sample size and test length, as these conditions are under the control of the researcher and the test builder.

## 2. METHOD

### *2.1 Data selection*

Students' answers to the PISA 2018 test's mathematics exam items served as the study's source of data. The original dataset consisted of 83 items and was obtained from the Human Resources Development Center, the organization responsible for conducting the test in Jordan. This central source provided the data file for the test. The present study examined the overall PISA assessment, followed by an analysis of the responses provided by students in Jordan. A study sample was subsequently selected, reflected in the dataset labelled Pocket No. 14, the total number of students there was 342.

To achieve the objectives of this study, three levels of test length were adopted: the first level consists of 30 randomly selected items, the second level consists of 20 randomly selected items from the first level items, and the third level consists of 10 randomly selected items from the second level items. Table 2 indicates the names of items and selected for each level of test length from Pocket No. 14 in the PISA mathematics test.

**Table 2: Items for each level of test length: 30, 20, 10**

| Item # | Item name | Test length 30 | Test length 20 | Test length 10 |
|---|---|---|---|---|
| $I_1$ | Cash Withdrawal - Q01 (Paper Scored Response) | × | × | × |
| $I_2$ | Cash Withdrawal - Q02 (Coded Paper Response) | × | | |
| $I_3$ | Tossing Coins - Q01 (Paper Scored Response) | × | × | × |
| $I_4$ | Containers - Q01 (Paper Scored Response) | × | × | × |
| $I_5$ | Running Tracks - Q01 (Coded Paper Response) | × | × | |
| $I_6$ | Number Check - Q01 [Part A] (Raw Paper Response) | × | | |
| $I_7$ | Number Check - Q01 [Part B] (Raw Paper Response) | × | | |
| $I_8$ | Number Check - Q01 (Paper Scored Response) | × | × | × |
| $I_9$ | Stop The Car - Q01 (Paper Scored Response) | × | × | |
| $I_{10}$ | Chair Lift - Q01 (Paper Scored Response) | × | | |
| $I_{11}$ | Chair Lift - Q02 (Paper Scored Response) | × | × | × |
| $I_{12}$ | Tile Arrangement - Q01 (Paper Scored Response) | × | × | |
| **Item #** | **Item name** | **Test length 30** | **Test length 20** | **Test length 10** |
| $I_{13}$ | Pipelines - Q01 (Paper Scored Response) | × | × | |
| $I_{14}$ | Lotteries - Q01 [Part A] (Raw Paper Response) | × | | |
| $I_{15}$ | Lotteries - Q01 [Part B] (Raw Paper Response) | × | × | × |
| $I_{16}$ | Lotteries - Q01 [Part C] (Raw Paper Response) | × | × | |
| $I_{17}$ | Lotteries - Q01 [Part D] (Raw Paper Response) | × | | |
| $I_{18}$ | Lotteries - Q01 (Paper Scored Response) | × | | |
| $I_{19}$ | Transport - Q01 [Part A] (Raw Paper Response) | × | | |
| $I_{20}$ | Transport - Q01 [Part B] (Raw Paper Response) | × | × | × |
| $I_{21}$ | Transport - Q01 [Part C] (Raw Paper Response) | × | × | × |
| $I_{22}$ | Transport - Q01 [Part D] (Raw Paper Response) | × | × | |
| $I_{23}$ | Transport - Q01 (Paper Scored Response) | × | × | |
| $I_{24}$ | Thermometer Cricket - Q01 (Coded Paper Response) | × | | |

| Item # | Item name | Test length 30 | Test length 20 | Test length 10 |
|---|---|---|---|---|
| $I_{25}$ | Thermometer Cricket - Q02 (Coded Paper Response) | × | × | |
| $I_{26}$ | Telephone Rates - Q01 (Paper Scored Response) | × | | |
| $I_{27}$ | Carbon Dioxide - Q01 (Coded Paper Response) | × | × | × |
| $I_{28}$ | Carbon Dioxide - Q03 (Coded Paper Response) | × | × | |
| $I_{29}$ | Fence - Q01 (Paper Scored Response) | × | × | |
| $I_{30}$ | Computer Game - Q01 (Paper Scored Response) | × | × | × |

Three different sample sizes were used: The first level is composed of the 324 students (male and female) who make up the entire sample of respondents for this booklet. The second level is composed of 200 students who were randomly picked from the sample, and the third level comprises 100 students who were also randomly selected.

***2.2 Data Analysis:***

DIF was detected using the MH-LOR method in the difference in sample size and test length by applying DDFS software. The results of the MH-LOR method were interpreted based on the following indicators:

**Z (LOR):** This indicator determines whether a DIF is present.

If the value of Z (LOR) > 2 or Z (LOR) < -2, this indicates the presence of DIF at the level of statistical significance ($\alpha = 0.05$).

**To determine the magnitude of DIF**:

If $|LOR| \geq 0.64$, the amount of DIF is significant.

If $0.43 \leq |LOR| < 0.64$, the amount of DIF is medium. If $|LOR| < 0.43$, the amount of DIF is small.

**SE (LOR):** Specifies the standard error value of LOR.

**BD. (Breslow-Day Chi-Square test):** This indicator was used to demonstrate the presence of non-uniform DIF for items at the level of statistical significance of 0.05, and its critical values were 3.84 (Penfield, 2010).

## 3. RESULTS / FINDINGS

The results of the detection of DIF items using the MH-LOR method were obtained by applying the DDFS program to the responses of individuals at the levels of sample sizes and test lengths, (9) runs (3 levels of sample size × 3 levels of test length) by the DDFS program were implemented to extract results.

In the beginning, the program was implemented three times, so that the sample size was fixed at its largest size, which is 342, with a change in the length of the test each time of implementation, so that we obtained DIF indicators for all items, and only indicators of DIF items were recorded in Table 3.

**Table 3: DIF items in the sample size of (342) at all levels of test length**

| Test Length | Item | LOR | SE (LOR) | Z (LOR) | BD | DIF magnitude | DIF group | DIF type |
|---|---|---|---|---|---|---|---|---|
| **30** | $I_{11}$ | 0.7913 | 0.2926 | 2.7044* | 0.708 | Significant | male | Uniform |
| | $I_{21}$ | -0.7208 | 0.2524 | -2.8558* | 0.007 | Significant | female | Uniform |
| | $I_{30}$ | -0.7286 | 0.2954 | -2.4665* | 0.123 | Significant | female | Uniform |
| **20** | $I_{11}$ | 0.7453 | 0.2866 | 2.6005* | 0.666 | Significant | male | Uniform |
| | $I_{21}$ | -0.8025 | 0.2503 | -3.2062* | 0.32 | Significant | female | Uniform |
| | $I_{27}$ | -1.341 | 0.5675 | -2.363* | 0.032 | Significant | female | Uniform |
| | $I_{30}$ | -0.7439 | 0.2752 | -2.7031* | 0 | Significant | female | Uniform |
| | $I_{1}$ | 0.6371 | 0.2938 | 2.1685* | 0.121 | Medium | male | Uniform |
| | $I_{11}$ | 0.865 | 0.2947 | 2.9352* | 1.687 | Significant | male | Uniform |

| Test Length | Item | LOR | SE (LOR) | Z (LOR) | BD | DIF magnitude | DIF group | DIF type |
|---|---|---|---|---|---|---|---|---|
| **10** | $I_{21}$ | -0.7715 | 0.2468 | -3.126* | 0.027 | Significant | female | Uniform |
| | $I_{27}$ | -1.4817 | 0.5761 | -2.5719* | 6.536 | Significant | female | Nonuniform |
| | $I_{30}$ | -0.8224 | 0.2872 | -2.8635* | 1.233 | Significant | female | Uniform |

By the same way, the program was re- implemented three more times by fixing the sample size at 200 and changing the test lengths, and the indicators of DIF items were recorded in Table 4.

**Table 4: DIF items in the sample size of (200) at all levels of test length**

| Test Length | Item | LOR | SE (LOR) | Z (LOR) | BD | DIF magnitude | DIF group | DIF type |
|---|---|---|---|---|---|---|---|---|
| **30** | $I_{30}$ | -0.9096 | 0.431 | -2.1104* | 0.108 | Significant | female | Uniform |
| **20** | $I_{21}$ | -0.9232 | 0.3608 | -2.5588* | 0.003 | Significant | female | Uniform |
| | $I_{30}$ | -0.99 | 0.3838 | -2.5795* | 0.029 | Significant | female | Uniform |
| **10** | $I_{1}$ | 0.7999 | 0.3581 | 2.2337* | 0.067 | Significant | male | Uniform |
| | $I_{21}$ | -0.7599 | 0.3227 | -2.3548* | 0.239 | Significant | female | Uniform |
| | $I_{30}$ | -1.4011 | 0.4468 | -3.1359* | 0.685 | Significant | female | Uniform |

Finally, the program was implemented three times to find the DIF indicators where the sample size was fixed at 100 and test lengths were changed, and the indicators of DIF items were recorded in Table 5.

**Table 5: DIF items in the sample size of (100) at all levels of test length**

| Test Length | Item | LOR | SE (LOR) | Z (LOR) | BD | DIF magnitude | DIF group | DIF type |
|---|---|---|---|---|---|---|---|---|
| **30** | $I_{11}$ | 1.6596 | 0.6444 | 2.5754* | 1.427 | Significant | male | Uniform |
| **20** | $I_{9}$ | 1.4224 | 0.5683 | 2.5029* | 4.14 | Significant | male | Nonuniform |
| **10** | $I_{11}$ | 2.3789 | 0.7323 | 3.2485* | 0.2 | Significant | male | Uniform |
| | $I_{27}$ | -1.8245 | 0.7694 | -2.3713* | 27.23 | Significant | female | Nonuniform |

Table 3 shows the numbers of DIF items, their size and type, and the associated DIF group when the largest sample size (all 342 students) is used for the test across different test length levels. The findings indicate that the MH method identified 3, 4, and 5 DIF items at 30, 20, and 10 test lengths, respectively. Notably, the number of DIF items increases as the test length decreases. When the sample size is reduced to 200, it's noticed that the number of DIF items decreases at all levels of the test length but still increases with shortening the test length; as shown in Table 4, the MH method was able to detect 1, 2, 3 DIF items at test lengths of 30, 20, and 10, respectively.

The number of DIF items seemed as low as possible when the sample size was lowered to the lowest threshold of 100. However, it also increased when the test length decreased, as Table 5 showed that the MH method revealed 1, 1, and 2 DIF items at test lengths 30, 20, and 10, respectively. The numbers of DIF items were summarized in all cases resulting from the intersection of levels of sample size and test length in Table 6

**Table 6: Numbers of DIF items at all levels of test length and sample size**

| Test length | | | | | |
|---|---|---|---|---|---|
| | | **30 Items** | **20 Items** | **Ten items** | **Total** |
| **Sample size** | **342** | 3 | 4 | 5 | 12 |
| | **200** | 1 | 2 | 3 | 6 |
| | **100** | 1 | 1 | 2 | 4 |
| | **Total** | 5 | 7 | 10 | 22 |

When glancing at the results presented in **Tables 3, 4, 5, and 6**, we find that when the test length is fixed for a certain length. The sample size is changed, the number of items that show DIF increases, and this means that the ability of the MH method for detecting DIF increases with increasing sample size.

## 4. DISCUSSION and CONCLUSION

### *4.1. Discussions*

Through an overview of the summary of the study results presented in Table 6, it is noted that the Mantel-Haenszel method was able to detect the highest number of DIF items at the highest sample size, which is 342 across all test lengths, and the number of DIF items decreases when the sample size is reduced, as is clear in the total column in Table 6, When looking at the total row in the same table, we note that the highest number of DIF items detected by the method was at the minimum test length of 10 items across all three sample sizes, and the number of DIF items increases when the length of the test is reduced.

The current finding aligns with the findings of (Arıkan et al., 2016), which showed that increasing the sample size improves the capacity and efficacy of the method in identifying DIF items. It also agrees with the result of (Gao, 2019) study, as well as the results of (Dorans & Holland, 1992) in that they suggested whenever feasible, the largest possible sample size for both focal and reference groups should be used in DIF. Based on the findings of this study, it can be concluded that to maximize the identification of DIF items, researchers should consider increasing the sample size applied.

Conversely, when the sample size remains constant and the test length changes by looking at the numbers of DIF items, it is observed that a decrease in test length corresponds to an increase in the number of DIF items. For instance, in the present study, the MH method exhibited the highest sensitivity in detecting DIF items compared to other test lengths, as it successfully identified 10 DIF items across all sample size levels at the lowest test length used.

When comparing sample size and test length, we can see that the best position was at a test length of 10 and a sample size of 342; the method was able to identify 5 DIF items at this length, and the least amount of DIF items appeared at a test length of 30 and a sample size of 100.

The issue of excluding DIF items is not at the expense of their quality; a good researcher balances statistical results and logical justification, so referring to the formulation, arbitration, treatment, and logical modification of those items is preferable. The items that show DIF for the gender variable were studied in more detail. The results showed that the DIF appeared in items (1, 8, and 11) in favor of males and items (21, 30, and 26) in favor of females.

Referring to those items' content, we see that the DIF items that favor males measure the domains of probability, geometry, and numbers. In contrast, DIF items in favour of females measure the areas of algebra, mathematical relations, and statistical interpretations, and this result agrees with the study by Taylor & Lee, 2012 (Taylor & Lee, 2012)

### 4.2. Conclusions

The study concluded that the MH method is affected by the difference in sample size and test length, as the method's ability to detect DIF items increased with the increase of sample size when fixation of test length at a specific limit. In contrast, it was observed that the method's ability decreases with increasing test length and fixing the sample size at a certain level. This study finding indicates that when the MH method is used for detecting DIF items, it is recommended to use a large sample size and a short test length to increase its effectiveness in detecting those items.

**REFERENCES**


Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29(1), 67-91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x

Alomari, H., Akour, M. M., & Al Ajlouni, J. (2023). The effect of Sample Size on Differential Item Functioning and Differential Distractor Functioning in multiple-choice items. *Psychology Hub*, 40(2), 17–24. https://doi.org/10.13133/2724-2943/17992

Arıkan, Ç., Uğurlu, S., & Atar, B. (2016). A DIF and bias study by using MIMIC, SIBTEST, Logistic Regression, and Mantel-Haenszel methods. *Hacettepe University Journal of Education*, 31(1), 34-52. DOI:10.16986/HUJE.2015014226

Camilli, G., Shepard, L. A., & Shepard, L. (1994). *Methods for identifying biased test items* (Vol. 4). SAGE: university of Michigan.

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization 1, 2. *ETS Research Report Series, 1992(1), i-40*. https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Eom, M. (2008). Underlying factors of MELAB listening construct. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, 6, 77–94.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied psychological measurement*, 29(4), 278-295. DOI:10.1177/0146621605275728

GAO, X. (2019). A comparison of six DIF detection methods. Unpublished Master Theses, University of Connecticut Graduate School, *https://digitalcommons.lib.uconn.edu/gs_theses/1411*

GU, K. (2023). Washback Effects of IELTS Test on Teachers' Adoption of Teaching Materials in the Classroom in China. *International Journal on Social & Education Sciences (IJonSES), 5(2).* DOI: https://doi.org/10.46328/ijonses.513

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. *ETS Research Report Series, 1986(*2), i-24. DOI: https://doi.org/10.1002/j.2330-8516.1986.tb00186.x

Ihlenfeldt, S. D., & Rios, J. A. (2023). A meta-analysis on the predictive validity of English language proficiency assessments for college admissions. *Language Testing*, 40(2), 276-299. DOI:10.1177/02655322221112364

Kabasakala, K., Arsan, N., Gok, B., & Kelecooglu, H. (2014). Comparing Performances (Type I error and Power) of IRT Likelihood Ratio SIBTEST and Mantel-Haenszel Methods in the Determination of Differential Item Functioning. *Educational Sciences: Theory & Practice*, 14(6), 2186-2193. DOI: 10.12738/estp.2014.6.2165

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

Marôco, J. (2021). Portugal: The PISA Effects on Education. In: Crato, N. (eds) Improving a Country's Education. Springer, Cham. https://doi.org/10.1007/978-3-030-59031-4_8

Mellenbergh, G. J. (1989). Item bias and item response theory. *International journal of educational research,* 13(2), 127-143.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied psychological measurement,* 17(4), 297-334. https://doi.org/10.1177/014662169301700401

Münch, R., & Wieczorek, O. (2023). Improving schooling through effective governance? The United States, Canada, South Korea, and Singapore are in the struggle for PISA scores. *Comparative Education,* 59(1), 59-76. DOI:10.1080/03050068.2022.2138176

Narayanon, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied psychological measurement,* 20(3), 257-274. https://doi.org/10.1177/014662169602000306

Park, G. (2008). Differential Item Functioning on an English Listening Test across Gender. *TESOL Quarterly*, 42(1), 115-123.

Penfield, R. D. (2010). DDFS: Differential distractor functioning software. *Applied psychological measurement,* 34(8), 646-647. https://doi.org/10.1177/0146621610375690

Penfield, R. D., & Camilli, G. (2006). Five Differential Item Functioning and Item Bias. *Handbook of statistics, 26, 125-167.* DOI:10.1016/S0169-7161(06)26005-X

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of*

*educational measurement*, 27(4), 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education*, 25(3), 246-280. https://doi.org/10.1080/08957347.2012.687650.

the MELAB Listening Test. Language Assessment Quarterly, 8, 361–385. DOI:10.1080/15434303.2011.628632

Vahid A., Christine C. & Lee O. (2011). An Investigation of Differential Item Functioning in

Wagner, A. (2004). A construct validation study of the extended listening sections of the ECRE and MELAB. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, 2, 1–23.

Wall, D., & Horák, T. (2006). The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 1, the baseline study. *ETS Research Report Series, 2006(1), i-199*. https://doi.org/10.1002/j.2333-8504.2006.tb02024.x

Wall, D., & Horák, T. (2008). The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 2, coping with change. *ETS Research Report Series, 2008(2), i-105.* https://doi.org/10.1002/j.2333-8504.2008.tb02123.x

Williams, S. (1997). The unbiased anchor bridging the gap between DIF and item bias. *Applied Measurement and Education*, 10(3), 253-267. https://doi.org/10.1207/s15324818ame1003_4