# SENSITIVITY OF DIFAS, jMetrik and STATA SOFTWARE IN DETERMINING GENDER DIF USING MH PROCEDURE


## Abstract

The analysis of differential item functioning (DIF) examines whether item responses differ according to background information such as gender, location, language spoken and type of school when people with equivalent ability levels have different probability of responding to items. This analysis can be performed by calculating various statistics, one of the most important being the Mantel-Haenszel, which can be carried out with software programs such as jMetrik, DIFAS and, more recently, STATA. In this perspective, the purpose of this study is to compare the sensitivity of these three software programs in detecting gender DIF and their corresponding effect size by using 50 multiple-choice integrated science real data. The procedural characteristics and the results obtained from the same dataset were thus compared by the three programs. DIFAS and jMetrik always provide equivalent results, while STATA is less accurate when using the thin matching strategy. The results also showed that DIFAS was the easiest to run, especially for testing practitioners, with the second offering a broader range of results for key characteristics for detecting DIF.


**Keywords:** STATA, jMetrik, DIFAS, Gender DIF, Effect size, MH Procedure

## Introduction

In the high-stakes testing world, assessment researchers consistently evaluate the fairness of tests and explore the reasons behind achievement gaps and poor academic performance of some examinees (Sabatini *et al.*, 2015; Huang *et al.*, 2016). The items of an examination show differential item functioning (DIF) when subjects with the same ability level for the characteristics or attributes being measured, but who belong to different groups (demographic, linguistic, or

cultural), have a different chance of giving a correct item response (Millsap & Everson, 1993). Differential item functioning (DIF) is a statistical characteristic of an item that shows the extent to which the item might be measuring different abilities for members of separate subgroups. Average item scores for subgroups having the same overall score on the test are compared to determine whether the item is measuring in essentially the same way for all subgroups (Camilli & Shepard, 1994). The presence of DIF requires review and judgment, and it does not necessarily indicate the presence of bias. DIF analysis indicates unexpected behaviour of items on a test. An item does not display DIF if people from different groups have a different probability to give a certain response; it displays DIF if and only if people from different groups *with the same underlying true ability* have a different possibility of giving a correct response.

DIF is usually studied by comparing two groups of subjects: the reference group (generally the majority/group having an advantage), and the focal group (generally a minority group/ group being at a disadvantage). Between these two groups, DIF may appear as uniform or non-uniform. In the former, there is no interaction between the score level of the attribute or trait being measured and membership of a given group, i.e. the probability of giving a correct response to the item is consistently higher for one group than the other across all score levels of the attribute. However, in the case of non-uniform DIF, there is an interaction, i.e. the likelihood of giving a correct response to the item in the two groups is not the same for all score levels of the attribute/trait/ability being measured (Chalmers, 2018).

Over the last decade, a lot of statistical methods developed for detecting measurement bias in psychological and educational tests have been reviewed. Earlier methods for assessing measurement bias generally have been replaced by more sophisticated statistical techniques, such as the Mantel-Haenszel procedure, the standardization approach, logistic regression models, and

item response theory approaches. There have been a wide variety of statistical techniques for evaluating DIF in both dichotomous and polytomous items (Hidalgo & Gomez-Benito, 2010; Potenza & Dorans, 1995). Among these, the Mantel-Haenszel (MH) statistic is regarded as a reference technique due to its ease of use and the fact that it can be applied to small samples. These characteristics have meant that numerous studies in the applied field, such as those conducted by the Educational Testing Service (ETS), have used the MH statistic to detect DIF. Its usefulness in the field measurement and evaluation has also been the emphasis of most research (Guilera, Gomez-Benito & Hidalgo, 2009).

Some specific software programs aimed at detecting DIF employing the MH procedure are now available, specifically, EZDIF (Waller, 1998), DIFAS (Penfield, 2005), EASY-DIF (Gonzalez, Padilla, Hidalgo, Gomez-Benito & Benitez, 2011), jMetrik (Meyer, 2009) and Stata (StataCorp, 1985). In this context, the purpose of this study was to examine the characteristics of three of these programs as well as their advantages and disadvantages. This was done by conducting a comparative analysis of a real science dataset using the three programs: DIFAS, jMetrik and Stata. This comparison was based not only on the instrumental and procedural characteristics of each software package but also on the results they provided following the analysis of common simulated and real datasets.

The Mantel-Haenszel (MH) statistical procedure (Mantel & Haenszel,1959) is subtle to only one type of differential item functioning (DIF). It is not designed to detect DIF that has a non-uniform effect across trait levels. By generalizing the model underlying the MH procedure, a more general DIF detection procedure has been developed (Swaminathan & Rogers, 1990). It consists of comparing the item performance of two groups (reference and focal), whose members were previously matched on the ability scale. The matching is done using the observed total test score

as a criterion or matching variable (Holland & Thayer, 1988). The Mantel-Haenszel statistic is based on a contingency table analysis. For dichotomous items, K contingency tables (2 × 2) are constructed for each item, where K is the number of test score levels into which the matching variable has been divided. Table 1 shows the 2 x 2 table for calculating the MH statistic for item *i* on a *j* score level in the test.

**Table 1: Score on $i^{th}$item in j score**

| Group | 1 | 0 | Total |
|---|---|---|---|
| Reference | *Aj* | *Bj* | $N_{Ri}$ |
| Focal | $C_j$ | $D_j$ | $N_{Fj}$ |
| Total | $N_{1j}$ | $N_{0j}$ | $N_j$ |

In typical applications of the MH procedure, an item shows uniform DIF if the odds of correctly answering the analysed item at a given score level *j* is different for the two groups at some level *j* of the matching variable. The odds ratio (α) is given by: α = (pRj/1- pRj)/ (pFj/1- pFj) in which pRj and pFj are the correct item response probabilities for the reference group and focal group, respectively. The test score level *j* is calculated as follows:

$p_{Fj} = \frac{C_j}{N_j}$ and $p_{Rj} = \frac{A_j}{N_{Rj}}$

The MH statistic for detecting DIF in an item is expressed as:

$$\frac{[|\sum_{j=1}^{K} A_j - \sum_{j=1}^{K} E(A_j)| - 0.5)]^2}{\sum_{j=1}^{K} Var(A_j)}$$

in which *E(Aj) = (NRjN1.j)/N..j and Var(Aj) = (NRjNFjN1.jN0.j)/(N..j)2(N..j – 1)*. The MH

statistic, under the null hypothesis, is distributed as a $\chi_2$ distribution with one degree of freedom. Under the MH procedure, an effect size estimate based on the common odds ratio $\alpha$ is expressed as

$$\alpha_{MH} = \frac{[\sum_{j=1}^{K} A_i D_i / N_{..j}}{\sum_{j=1}^{K} B_i C_i / N_{..j}}$$

Holland and Thayer (1988) proposed a logarithmic transformation of α for interpretive purposes, intending to obtain a symmetrical scale in which a zero value indicates an absence of DIF, a negative value indicates that the item favours the reference group over the focal group, and a positive value indicates DIF in the opposite direction. This transformation is expressed as

$\Delta\alpha_{MH} = -2.35 \ln (\alpha_{MH})$

Based on this transformation, Zwick and Ercikan (1989) proposed the following interpretation guidelines to evaluate the DIF effect size:

1. Type A items−negligible DIF: items with $\Delta\alpha_{MH} < |1|$.
2. Type B items−moderate DIF: items with $|1| \leq \_ \Delta\alpha_{MH} \leq |1.5|$, and the MH test statistically significant.
3. Type C items−large DIF: items with $\Delta\alpha_{MH} > |1.5|$, and the MH test statistically significant.

Zwick and Ercikan (1989) pointed out that Type B items could be used in the test if there are no others to replace them, and that Type C items will be selected only if they are necessary to meet test specifications. Based on the general characteristics of the MH procedure, new statistics have also been developed, for example, the Breslow-Day chi-square (Breslow & Day, 1980) and new procedures for DIF detection such as the combined decision rule (Penfield, 2003).

**Methodology**

A real dataset was used to compare the three software programs. The dataset comes from the responses 2016 Integrated Science West African Senior Secondary Certificate Examination (WASSCE) which were organized by West African Examination Council (WAEC) to assess candidates who have completed three-year Senior High School Education. The WASSCE Integrated Science consists of three papers, Paper 1, Paper 2 and Paper 3. Paper 1 will usually be an objective – type test paper. Paper 2 will consist of structured questions or essay questions, essentially testing, Application of Knowledge", but also consisting of some questions on Knowledge and Understanding. Paper 3 will be the practical test paper which is intended to evaluate candidates' knowledge and ability in the diversity of matter, cycles, systems, energy and interactions of matter themes. Paper 1 used in the current study consist of 50 multiple-choice items with four response categories that can be coded according to a dichotomous system.

For this study, a sample of 10, 313 respondents made up of 5,125 males and 5,118 females were extracted from the survey database In the DIF analysis the male group was defined as the reference group and the female examinees as the focal group. The participants' responses were coded dichotomously. The analysis of characteristics and the comparison of the DIFAS, jMetrik, and Stata programs took into account several aspects. The first of these concerned procedural parameters such as how the programs could be obtained (availability, material, etc.), data handling and the analyses possible in each case. Subsequently, the results they provided were compared by analyzing a common dataset.

*Availability*

DIFAS, jMetik and Stata are currently the most used free software programs for evaluating DIF. These programs are run in Windows. All three can be obtained by contacting the authors and

downloading from the software web. The program installer for each package comes with a user manual.

*DIF Analyses for Dichotomous Items*

Firstly, procedural aspects related to the formatting and input of datasets for each of the three programs are considered. The steps required before analysis are described in each case.

jMetrik:

There are many ways to conduct a differential item functioning (DIF) analysis. jMetrik provides several useful statistics including the Mantel-Haenszel chi-square procedure, commons odds ratio effect size, standardized p-dif effect size, and ETS DIF classification levels. These statistics allow you to judge the statistical and practical significance of DIF. You must complete item scoring before you conduct a DIF analysis. To conduct a DIF analysis you need a matching variable such as a sum score. If one does not exist in your data table, you must create it using the Test Scaling procedures and possibly the Ranking procedures.

1. Create a matching variable: If a matching variable does not exist in your data you can create one by computing a sum score. See the instructions for Test Scaling for directions on how to create a sum score. If a matching variable exists, then use it in the next step.
2. Choose thick or thin matching: Thin matching involves all levels of a sum score. To use thin matching, select a sum score variable as your matching variable in DIF analysis. This method provides the best control over the measured trait, but it may result in sparse tables and omitted responses. Think matching preserves more of the data but gives you less control over the measured trait. For think matching group examinees into ordered groups such as deciles. Use the Deciles option of the Ranking procedure to rank examinees into ten groups. Use the new decile variable as your matching variable in the DIF analysis.

3. Click Analyze > DIF: Mantel-Haenszel to start the DIF analysis dialog.
4. Select the items you would like to study and move them to the top-right list by clicking the first select button.
5. Select the matching variable and move it to the *Matching Variable field* by clicking the second select button.
6. Select the DIF grouping variable and move it to the *Group By field* by clicking the last select button. An example grouping variable is a gender.
7. Identify the Focal and reference group codes that are in your grouping variable. For example, the code F might indicate females in your DIF group variable and the code M might represent males. The case of the focal and reference group codes must match the case of the values listed for the DIF group variable. IF you use the wrong case, the program will not recognize the values in the group variable.
8. You can run the analysis at this point, but you may want to change some of the default options. *Binary item effect size* – The default value is *Common odds ratio option*. This statistic ranges from 0 to positive infinity and has an expected value of unity. To use a more symmetric effect size, choose the *ETS Delta option*. The ETS statistic is a transformation of the common odds ratio that has values that range from about -4 to +4 and are centred about zero.

DIFAS: To run DIF analyses for dichotomous items, follow these steps:

1. In the “Select Items” list (the list at the far left of the Dichotomous Models Window), select the items to be studied for DIF. Note that by default, the “Range of Items” option is selected. In this mode, a range of items can be selected by first selecting the lower item

of the range and then the upper item of the range. To turn off the "Range of Items" option, click on the "Individual Items" option.

2. In the "Select Groups" list (the list in the middle of the Dichotomous Models Window), select the variable designating the groups of interest. DIFAS permits only the comparison of two groups at a time. Below the variable selection box, specify the value designating reference group members, and the value designating focal group members. Any two numeric values may be used to designate the reference and focal groups.
3. In the "Select Stratify" list (the list at the far right of the Dichotomous Models Window), specify the stratifying (or matching) variable. If you wish the total test score obtained from the items selected to be included in the DIF analysis to serve as the stratifying variable, ensure that the option "Stratify by sum" is selected. In this case, no variable needs to be selected from the "Select Stratify" list. Note that when DIFAS computes the total test score, any individual having a missing value is assigned a missing value for the total test score, and thus is ignored in the analyses. Thus, if you wish a total test score to be computed for individuals having missing values, be sure to code all missing values as "0" before importing the data file into DIFAS. If you wish the stratifying variable to be some other variable in the data file, click the option "Stratify by external", and then select the external variable from the "Select Stratify" list.
4. By default, DIFAS uses a stratum size of 1. You can, however, change this size by specifying a stratum size in the box located in the lower right portion of the Dichotomous Models Window. For example, if a stratum width of 2 is desired, then the user would change the 1 to a 2 in the "Stratum Size" box.

5. DIFAS can print out the number of reference and focal group members located in each stratum of the stratifying variable. To do so, select the "Print strata n" check box in the lower right portion of the Dichotomous Models Window. Once these three steps are completed, click on the OK button. The resulting output consists of two tables – a table containing the relevant DIF statistics, and a table containing the conditional differences in mean item score between the reference and focal groups at ten intervals across the stratifying variable continuum (the lower and upper limits of each of the ten intervals are presented at the top of each interval).

STATA: Stata can import data in a variety of formats:

This includes ASCII data formats (such as CSV or databank formats) and spreadsheet formats (including various Excel formats). Stata's proprietary file formats have changed over time, although not every Stata release includes a new dataset format. Every version of Stata can read all older dataset formats, and can write both the current and most recent previous dataset format, using the same old command. Thus, the current Stata release can always open datasets that were created with older versions, but older versions cannot read newer format datasets. In analysing DIF with Mantel Haenszel in STATA, difmh calculates the Mantel–Haenszel (MH) $\chi^2$ and common odds ratio for dichotomously scored items. The MH statistics are used to determine whether an item exhibit uniform differential item functioning (DIF) between two observed groups, that is, whether an item favours one group relative to the other for all values of the latent trait. To illustrate the MH DIF analysis, STATA uses students' responses that are coded 1 for correct and 0 for incorrect and with code 1 for the focal group and 0 for the reference group. The output usually contains the Chi$^2$ and Prob. Columns which contain the MH $\chi^2$ statistic with the associated significance

level. Items exhibit DIF based on a 5% significance level. However, significant statistics do not tell us anything about the amount or direction of DIF exhibited by an item. The last three columns present the MH common odds ratio with the associated confidence interval. A common odds ratio greater than 1 indicates DIF in favour of the focal group. A visual examination of the output table sometimes becomes cumbersome even for a moderate number of items. It is therefore advisable to ask difmh to display only items whose p-value falls below a certain significance level with the maxp (.05) option.

**Identifying DIF**

jMetrik: According to Meyer, identifying DIF involves a combination of statistical and practical significance.

*Statistical Measure*

**jMetrik**: It uses Cochran-Mantel-Haenszel (CMH) Chi-square Statistic which is a chi-square statistic that compares the proportions correct across the matching subgroups of persons of the same ability. The null hypothesis is that there is no DIF and the expected value of CMH chi-square is 1. Alternatively, we assume there is DIF and the expected value of CMH is greater 1. The value 3.84 is the .05 critical value. The test is usually two-tailed. One must know that the value of the statistic indicates only a difference, not the direction of the difference.

**DIFAS**: It provides Mantel-Haenszel Chi-Square (MH CHI) which is distributed as chi-square with one degree of freedom. Critical values of this statistic are 3.84 for a Type I error rate of 0.05 and 6.63 for a Type I error rate of 0.01, Mantel-Haenszel Common Log-Odds Ratio (MH LOR) is asymptotically normally distributed, Standard Error of the Mantel-Haenszel Common Log-Odds Ratio (LOR SE) here is the nonsymmetric estimator, Standardized Mantel-Haenszel Log-

Breslow-Day Chi-Square (BD**)** which test of the trend in odds ratio heterogeneity is distributed as chi-square with one degree of freedom. Critical values of this statistic are 3.84 for a Type I error rate of 0.05 and 6.63 for a Type I error rate of 0.01. This statistic is effective

at detecting non-uniform DIF, Combined Decision Rule (CDR) flags any item for which either the Mantel-Haenszel chi-square or the Breslow-Day chi-square statistic is significant at a Type I error rate of 0.025 (Penfield, 2003). The message OK is printed if neither statistic is significant, and the message FLAG is printed if either statistic is significant.

**STATA:** The statistic is evaluated against a $\chi^2$ distribution with one degree of freedom. A significant MH $\chi^2$ statistic suggests the presence of DIF in an item, however, the statistic does not indicate the amount of DIF.

*Practical Measures*

jMetrick/ STATA: A measure of effect size is determined by the ratio of odds of a person in the reference group getting an item correct to the odds of a person in the focal group getting the item correct. The rules are No DIF: Odds ratio is 1, DIF favouring the reference group: > 1 and DIF favouring the focal group: < 1.

DIFAS: A measure of effect size is determined by the log-odds ratio (LOR Z) –divided by the estimated standard error. A value greater than 2.0 or less than –2.0 may be considered evidence of the presence of DIF.

*ETS delta statistic*

For jMetrik, a transformation of the odds-ratio into a Z-like statistic where no DIF is ETS = 0, DIF favouring the reference group as ETS > 0, and DIF favouring the focal group as ETS < 0 are used whiles DIFAS uses MH LOR positive values to indicate DIF in favour of the reference group, and negative values indicate DIF in favour of the focal groups.

*Classification of items.*

jMetrik: Meyer describes the following classification scheme. Each item is classified as either A, B, or C. A is seen as a good item and Cochran-Mantel-Haenszel (CMH) Chi-square Statistic is < 3.84, so p-value > .05. Common odds ratio between 0.65 and 1.53 are usually used. The B is classified as a questionable item whiles C is seen as a poor item. Common odds ratio < 0.53 and upper bound of CI < .65, i.e., both the item and the confidence interval have to be very low or common odds ratio > 1.89 and lower bound of CI > 1.53, i.e., both the item and the confidence interval have to be very high. For items with the potential DIF (B or C), the jMetrik program puts after the letter (+) if it favours the focal group (-) if it favours the reference group. The jMetrik program automatically applies the rules in its output tables. Meyer (2014, p. 75) says, "' C' items have large amounts of DIF and are of concern. In most cases, they should be eliminated from a test. The only reason to include a 'C' item is if its elimination leads to a problem with content validity that is more severe than the problem caused by DIF."

DIFAS uses the ETS categorization scheme by (Zieky, 1993) to categorizes items as having small (A), moderate (B), and large (C) levels of DIF.

STATA: This transformation is expressed as $\Delta_{\alpha MH}$= –2.35 ln ($\alpha$) where $\alpha$ is the value of the odds ratio, Zwick and Ercikan (1989) proposed the following interpretative rules to evaluate the DIF effect size: 1. type A items/negligible DIF: these are items in which $\Delta_{\alpha MH}$<|1|; 2. type B items/moderate DIF: items in which |1| ≤ $\Delta_{\alpha MH}$ ≤|1.5| and where MH proved to be statistically significant; and 3. type C items/large DIF: items in which DaMH > |1.5|

and MH proved to be statistically significant. Zwick and Ercikan (1989) point out those type B items can be used in the test if there are no others to replace them, whereas type C items, will only be selected if they are necessary to achieve test specifications.

**Comparison of results**

This section presents the results of the comparison between the three programs. The jMetrik, DIFAS, and STATA used the same matching strategy (Thin matching), thus jMetrik used thin matching which involves all levels of a sum score as well as Stata whiles DIFAS uses the total test score obtained from the items selected to be included in the DIF analysis to serve as the stratifying variable (matching variable). The jMetrik output includes the item numbers, chi-square value, p-value, valid number, Effect size at 95% confidence interval values and the class, thus the level of DIF effect. STATA, on the other hand, provides the researcher with the item number, chi-square value, odd ratio value and 95% confidence interval values whiles the DIFAS output includes the item name, Mantel-Haenszel chi-square(MH CHI), Mantel-Haenszel common log-odds ratio and estimated standard error(MH LOR & LOR SE), Standardized Mantel-Haenszel common log-odds ratio(LOR Z), Breslow-Day test of the trend in odds ratio heterogeneity (BD), Combined Decision Rule (CDR) and Educational Testing Service (ETS) classification scheme.

The results of the analyses will focus on the number of items identified as DIF, the number of items that exhibited DIF in favour of male and female students. It will also look at the effect size of the items that exhibited DIF.

| Item | Chi-square | p-value | Valid N | E.S. (95% C.I.) | | | Class |
|---|---|---|---|---|---|---|---|
| q1 | 798.45 | 0.00 | 7396 | 6.01 ( | 6.49, | 5.52) | C+ |
| q2 | 0.96 | 0.33 | 2504 | 0.19 ( | 0.56, | -0.19) | A |
| q3 | 1.17 | 0.28 | 4932 | 0.31 ( | 0.89, | -0.27) | A |
| q4 | 82.74 | 0.00 | 2428 | 1.80 ( | 2.20, | 1.41) | C+ |
| q5 | 29.00 | 0.00 | 4932 | -1.02 ( | -0.64, | -1.39) | B- |
| q6 | 29.33 | 0.00 | 2504 | -1.36 ( | -0.86, | -1.86) | B- |
| q7 | 20.54 | 0.00 | 2504 | -0.87 ( | -0.49, | -1.24) | A |
| q8 | 13.59 | 0.00 | 4892 | 0.84 ( | 1.29, | 0.39) | A |
| q9 | 9.39 | 0.00 | 4932 | 0.84 ( | 1.38, | 0.31) | A |
| q10 | 15.71 | 0.00 | 4932 | -0.71 ( | -0.36, | -1.07) | A |
| q11 | 67.97 | 0.00 | 4932 | -1.58 ( | -1.20, | -1.96) | C- |
| q12 | 125.88 | 0.00 | 4932 | 2.14 ( | 2.52, | 1.76) | C+ |
| q13 | 3.69 | 0.05 | 4932 | -0.33 ( | 0.01, | -0.67) | A |
| q14 | 0.00 | 0.95 | 4932 | -0.01 ( | 0.33, | -0.35) | A |
| q15 | 231.67 | 0.00 | 4932 | 3.15 ( | 3.56, | 2.73) | C+ |
| q16 | 22.03 | 0.00 | 7809 | -0.64 ( | -0.37, | -0.91) | A |
| q17 | 5.36 | 0.02 | 7809 | 0.30 ( | 0.55, | 0.05) | A |
| q18 | 7.23 | 0.01 | 4932 | 0.61 ( | 1.05, | 0.16) | A |
| q19 | 211.99 | 0.00 | 4932 | 9.09 ( | 11.18, | 7.01) | C+ |
| q20 | 0.04 | 0.85 | 4932 | 0.04 ( | 0.42, | -0.34) | A |
| q21 | 7.62 | 0.01 | 2504 | 0.56 ( | 0.96, | 0.16) | A |
| q22 | 32.38 | 0.00 | 4932 | 1.18 ( | 1.59, | 0.77) | B+ |
| q23 | 0.26 | 0.61 | 4892 | -0.10 ( | 0.28, | -0.49) | A |
| q24 | 24.42 | 0.00 | 2428 | 1.05 ( | 1.46, | 0.63) | B+ |
| q25 | 42.36 | 0.00 | 2428 | 1.26 ( | 1.64, | 0.88) | B+ |
| q26 | 172.25 | 0.00 | 7396 | -2.24 ( | -1.89, | -2.58) | C- |
| q27 | 88.65 | 0.00 | 7396 | -1.57 ( | -1.24, | -1.91) | C- |
| q28 | 4.06 | 0.04 | 4932 | 0.58 ( | 1.14, | 0.01) | A |
| q29 | 57.21 | 0.00 | 4892 | 1.62 ( | 2.05, | 1.19) | C+ |
| q30 | 10.94 | 0.00 | 4892 | -0.62 ( | -0.25, | -0.99) | A |
| q31 | 16.53 | 0.00 | 7396 | -0.59 ( | -0.31, | -0.88) | A |
| q32 | 30.82 | 0.00 | 4932 | -1.40 ( | -0.89, | -1.91) | B- |
| q33 | 20.22 | 0.00 | 4892 | 0.87 ( | 1.25, | 0.49) | A |
| q34 | 58.62 | 0.00 | 4932 | 1.98 ( | 2.51, | 1.44) | C+ |
| q35 | 47.32 | 0.00 | 4932 | 1.29 ( | 1.65, | 0.92) | B+ |
| q36 | 1.02 | 0.31 | 4892 | -0.20 ( | 0.19, | -0.59) | A |
| q37 | 6.05 | 0.01 | 7396 | 0.36 ( | 0.65, | 0.07) | A |
| q38 | 27.70 | 0.00 | 4932 | -1.28 ( | -0.80, | -1.76) | B- |
| q39 | 38.06 | 0.00 | 2504 | -1.23 ( | -0.83, | -1.62) | B- |
| q40 | 0.92 | 0.34 | 2504 | -0.18 ( | 0.19, | -0.56) | A |
| q41 | 40.83 | 0.00 | 4932 | -1.31 ( | -0.90, | -1.72) | B- |
| q42 | 0.29 | 0.59 | 2504 | 0.10 ( | 0.48, | -0.27) | A |
| q43 | 32.33 | 0.00 | 4932 | 1.02 ( | 1.37, | 0.67) | B+ |
| q44 | 1.21 | 0.27 | 4932 | -0.25 ( | 0.20, | -0.70) | A |
| q45 | 117.79 | 0.00 | 4932 | -2.33 ( | -1.90, | -2.77) | C- |
| q46 | 132.09 | 0.00 | 2428 | 2.29 ( | 2.69, | 1.90) | C+ |
| q47 | 14.79 | 0.00 | 4892 | -0.64 ( | -0.31, | -0.97) | A |
| q48 | 40.77 | 0.00 | 7396 | 0.91 ( | 1.19, | 0.63) | A |
| q49 | 36.64 | 0.00 | 4932 | -1.26 ( | -0.85, | -1.68) | B- |
| q50 | 27.03 | 0.00 | 7396 | -0.84 ( | -0.52, | -1.16) | A |

*Figure 1*: Data output using jMetrik

**Table 2: Results of Gender DIF using jMetrik**

| | **Effect Size** | | | |
|---|---|---|---|---|
| **Gender** | **A** | **B** | **C** | **Total** |
| **Male** | 0 | 7 | 4 | **11** |
| **Female** | 17 | 5 | 8 | **30** |

| | | | | |
|---|---|---|---|---|
| **Total** | **17** | **12** | **12** | **41** |

Results from Figure 1 and Table 2 indicate that even though thirty (30) items exhibited DIF in favour of females, most of these items (17) exhibited negligible DIF(A). Also, an equal number (12) of items exhibited B and C levels of DIF. The items that exhibited B and C levels of DIF give indications that the items contain a moderate and large amount of DIF. Meyer (2014, p. 75) says, C' items have large amounts of DIF and are of concern. In most cases, they should be eliminated from a test. The only reason to include a 'C' item is if its elimination leads to a problem with content validity that is more severe than the problem caused by DIF."

DIF STATISTICS: DICHOTOMOUS ITEMS

| Name | MH CHI | MH LOR | LOR SE | LOR Z | BD | CDR | ETS |
|---|---|---|---|---|---|---|---|
| Var 2 | 896.3485 | -4.4325 | 0.2578 | -17.1936 | 69.594 | Flag | C |
| Var 3 | 17.485 | -0.6883 | 0.1767 | -3.8953 | 32.62 | Flag | B |
| Var 4 | 0.002 | -0.0148 | 0.1367 | -0.1083 | 110.124 | Flag | A |
| Var 5 | 60.8141 | -0.973 | 0.1355 | -7.1808 | 13.299 | Flag | C |
| Var 6 | 61.5291 | 1.0442 | 0.1501 | 6.9567 | 72.488 | Flag | C |
| Var 7 | 29.5773 | 0.9585 | 0.1876 | 5.1093 | 4.731 | Flag | C |
| Var 8 | 59.1858 | 1.7152 | 0.245 | 7.0008 | 21.498 | Flag | C |
| Var 9 | 0.056 | -0.0345 | 0.121 | -0.2851 | 52.54 | Flag | A |
| Var 10 | 28.084 | -0.7378 | 0.1496 | -4.9318 | 2.27 | Flag | C |
| Var 11 | 48.6425 | 0.7055 | 0.1031 | 6.8429 | 1.84 | Flag | C |
| Var 12 | 186.4929 | 1.8927 | 0.153 | 12.3706 | 7.112 | Flag | C |
| Var 13 | 152.5295 | -1.6336 | 0.1437 | -11.3681 | 2.285 | Flag | C |
| Var 14 | 10.2051 | 0.3486 | 0.1101 | 3.1662 | 3.514 | Flag | A |
| Var 15 | 9.0186 | 0.2528 | 0.0902 | 2.8027 | 213.785 | Flag | A |
| Var 16 | 286.794 | -2.8803 | 0.2078 | -13.8609 | 2.8 | Flag | C |
| Var 17 | 11.2418 | 0.2324 | 0.0691 | 3.3632 | 1.333 | Flag | A |
| Var 18 | 2.5199 | -0.0986 | 0.0621 | -1.5878 | 6.899 | Flag | A |
| Var 19 | 13.732 | -0.4393 | 0.1229 | -3.5745 | 17.311 | Flag | B |
| Var 20 | 234.9374 | -4.2298 | 0.4662 | -9.0729 | 74.395 | Flag | C |
| Var 21 | 0.1302 | 0.0605 | 0.1461 | 0.4141 | 8.71 | Flag | A |
| Var 22 | 16.1524 | -0.6233 | 0.1609 | -3.8738 | 5.089 | Flag | B |
| Var 23 | 19.7402 | -0.5196 | 0.1187 | -4.3774 | 0.242 | Flag | B |
| Var 24 | 29.31 | 0.6825 | 0.1262 | 5.4081 | 12.612 | Flag | C |
| Var 25 | 4.2895 | -0.2907 | 0.1375 | -2.1142 | 7.202 | Flag | A |
| Var 26 | 30.2199 | -0.6253 | 0.1192 | -5.2458 | 0.033 | Flag | B |
| Var 27 | 211.9498 | 1.2268 | 0.0894 | 13.7226 | 0.845 | Flag | C |
| Var 28 | 81.3936 | 0.7679 | 0.0883 | 8.6965 | 0.072 | Flag | C |
| Var 29 | 4.0237 | -0.3233 | 0.1597 | -2.0244 | 64.941 | Flag | A |
| Var 30 | 50.359 | -0.6687 | 0.0994 | -6.7274 | 104.336 | Flag | C |
| Var 31 | 14.5634 | 0.3324 | 0.0913 | 3.6407 | 40.427 | Flag | A |
| Var 32 | 21.9884 | 0.378 | 0.0826 | 4.5763 | 62.392 | Flag | A |
| Var 33 | 42.9648 | 0.7648 | 0.1263 | 6.0554 | 174.274 | Flag | C |
| Var 34 | 3.035 | -0.1839 | 0.1089 | -1.6887 | 8.421 | Flag | A |
| Var 35 | 42.3428 | -0.7957 | 0.134 | -5.9381 | 48.554 | Flag | C |
| Var 36 | 13.5384 | -0.3753 | 0.1043 | -3.5983 | 18.606 | Flag | A |
| Var 37 | 0.1409 | 0.0483 | 0.1165 | 0.4146 | 34.572 | Flag | A |
| Var 38 | 10.4339 | -0.2593 | 0.0814 | -3.1855 | 3.082 | Flag | A |
| Var 39 | 17.6586 | 0.5663 | 0.1426 | 3.9712 | 82.166 | Flag | B |
| Var 40 | 74.3356 | 1.142 | 0.142 | 8.0423 | 0.007 | Flag | C |
| Var 41 | 0.0007 | 0.0209 | 0.1805 | 0.1158 | 4.422 | OK | A |
| Var 42 | 56.6509 | 0.7389 | 0.1042 | 7.0912 | 108.151 | Flag | C |
| Var 43 | 0.1167 | -0.0684 | 0.1648 | -0.415 | 1.895 | OK | A |
| Var 44 | 9.4774 | -0.3146 | 0.1028 | -3.0603 | 3.421 | Flag | A |
| Var 45 | 0.0059 | -0.0143 | 0.1101 | -0.1299 | 4.455 | OK | A |
| Var 46 | 194.2749 | 1.5495 | 0.1226 | 12.6387 | 48.166 | Flag | C |
| Var 47 | 135.2251 | -1.4562 | 0.1321 | -11.0235 | 30.991 | Flag | C |
| Var 48 | 22.4962 | 0.3803 | 0.0813 | 4.6777 | 36.583 | Flag | A |
| Var 49 | 55.7226 | -0.5263 | 0.0727 | -7.2393 | 42.075 | Flag | B |
| Var 50 | 66.8015 | 0.849 | 0.1094 | 7.7605 | 49.455 | Flag | C |
| Var 51 | 34.3577 | 0.4368 | 0.0756 | 5.7778 | 23.721 | Flag | B |

*Figure 2*: Data output using DIFAS

**Table 3: Results of Gender DIF using DIFAS**

| Gender | Effect Size | | | Total |
|---|---|---|---|---|
| | A | B | C | |

| | | | | |
|---|---|---|---|---|
| **Male** | 8 | 2 | 13 | **23** |
| **Female** | 9 | 6 | 9 | **24** |
| **Total** | **17** | **8** | **22** | **47** |

Figure 2 and Table 3 show that 47 out of 50 items exhibited DIF. Out of the 47 items, 24 items favoured females while 23 showed DIF in favour of male students. According to Longford, Holland and Thayer (1993) if items are classified as A one can still include the item. If the item is classified as level-B one should examine if there are other items one can choose to include in the test instead. Finally, an item classified as C should only be chosen if it meets essential specifications but documentation and corroboration by a reviewer are required.

Mantel-Haenszel DIF Analysis

| Item | Chi2 | Prob. | Odds Ratio | [95% Conf. | Interval] |
|---|---|---|---|---|---|
| Q1 | 896.35 | 0.0000 | 84.1417 | 50.7613 | 139.4728 |
| Q2 | 17.48 | 0.0000 | 1.9904 | 1.4079 | 2.8139 |
| Q3 | 0.00 | 0.9647 | 1.0149 | 0.7763 | 1.3268 |
| Q4 | 60.81 | 0.0000 | 2.6458 | 2.0289 | 3.4503 |
| Q5 | 61.53 | 0.0000 | 0.3520 | 0.2623 | 0.4723 |
| Q6 | 29.58 | 0.0000 | 0.3835 | 0.2655 | 0.5538 |
| Q7 | 59.19 | 0.0000 | 0.1799 | 0.1113 | 0.2908 |
| Q8 | 0.06 | 0.8129 | 1.0351 | 0.8165 | 1.3121 |
| Q9 | 28.08 | 0.0000 | 2.0914 | 1.5598 | 2.8041 |
| Q10 | 48.64 | 0.0000 | 0.4939 | 0.4035 | 0.6045 |
| Q11 | 186.49 | 0.0000 | 0.1507 | 0.1116 | 0.2034 |
| Q12 | 152.53 | 0.0000 | 5.1224 | 3.8652 | 6.7883 |
| Q13 | 10.21 | 0.0014 | 0.7057 | 0.5688 | 0.8756 |
| Q14 | 9.02 | 0.0027 | 0.7766 | 0.6508 | 0.9267 |
| Q15 | 286.79 | 0.0000 | 17.8196 | 11.8589 | 26.7763 |
| Q16 | 11.24 | 0.0008 | 0.7926 | 0.6923 | 0.9076 |
| Q17 | 2.52 | 0.1124 | 1.1036 | 0.9771 | 1.2465 |
| Q18 | 13.73 | 0.0002 | 1.5516 | 1.2194 | 1.9743 |
| Q19 | 234.94 | 0.0000 | 68.7018 | 27.5494 | 171.3264 |
| Q20 | 0.13 | 0.7182 | 0.9413 | 0.7069 | 1.2534 |
| Q21 | 16.15 | 0.0001 | 1.8651 | 1.3606 | 2.5568 |
| Q22 | 19.74 | 0.0000 | 1.6814 | 1.3325 | 2.1217 |
| Q23 | 29.31 | 0.0000 | 0.5054 | 0.3946 | 0.6472 |
| Q24 | 4.29 | 0.0383 | 1.3374 | 1.0215 | 1.7509 |
| Q25 | 30.22 | 0.0000 | 1.8689 | 1.4796 | 2.3606 |
| Q26 | 211.95 | 0.0000 | 0.2932 | 0.2461 | 0.3494 |
| Q27 | 81.39 | 0.0000 | 0.4640 | 0.3902 | 0.5517 |
| Q28 | 4.02 | 0.0449 | 1.3817 | 1.0104 | 1.8894 |
| Q29 | 50.36 | 0.0000 | 1.9517 | 1.6063 | 2.3713 |
| Q30 | 14.56 | 0.0001 | 0.7172 | 0.5997 | 0.8578 |
| Q31 | 21.99 | 0.0000 | 0.6852 | 0.5827 | 0.8057 |
| Q32 | 42.96 | 0.0000 | 0.4654 | 0.3634 | 0.5961 |
| Q33 | 3.03 | 0.0815 | 1.2019 | 0.9710 | 1.4877 |
| Q34 | 42.34 | 0.0000 | 2.2159 | 1.7040 | 2.8816 |
| Q35 | 13.54 | 0.0002 | 1.4554 | 1.1864 | 1.7854 |
| Q36 | 0.14 | 0.7074 | 0.9529 | 0.7583 | 1.1973 |
| Q37 | 10.43 | 0.0012 | 1.2960 | 1.1049 | 1.5201 |
| Q38 | 17.66 | 0.0000 | 0.5676 | 0.4293 | 0.7506 |
| Q39 | 74.34 | 0.0000 | 0.3192 | 0.2417 | 0.4216 |
| Q40 | 0.00 | 0.9794 | 0.9794 | 0.6875 | 1.3951 |
| Q41 | 56.65 | 0.0000 | 0.4776 | 0.3894 | 0.5859 |
| Q42 | 0.12 | 0.7327 | 1.0708 | 0.7752 | 1.4791 |
| Q43 | 9.48 | 0.0021 | 1.3697 | 1.1198 | 1.6754 |
| Q44 | 0.01 | 0.9386 | 1.0144 | 0.8175 | 1.2586 |
| Q45 | 194.27 | 0.0000 | 0.2123 | 0.1670 | 0.2700 |
| Q46 | 135.23 | 0.0000 | 4.2897 | 3.3109 | 5.5579 |
| Q47 | 22.50 | 0.0000 | 0.6836 | 0.5830 | 0.8017 |
| Q48 | 55.72 | 0.0000 | 1.6926 | 1.4679 | 1.9517 |
| Q49 | 66.80 | 0.0000 | 0.4278 | 0.3452 | 0.5302 |
| Q50 | 34.36 | 0.0000 | 0.6461 | 0.5571 | 0.7492 |

*Figure 2*: Data output using STATA

**Table 4: Results of Gender DIF using STATA**

| Gender | Effect Size | Total |
|---|---|---|

| | A | B | C | |
|---|---|---|---|---|
| **Male** | 6 | 3 | 9 | **21** |
| **Female** | 6 | 6 | 12 | **21** |
| **Total** | **12** | **9** | **21** | **42** |

Results from STATA as shown in Figure 2 and Table 4 indicate that 42 out of 50 items exhibited gender DIF. Out of the 42 items that showed DIF, 12 were at the A level and therefore can be ignored and be used in assessing students. The results also indicated that 21 items out of the 42 were at the C-level which means these items cannot be used to assess students since they contain large DIF.

**Table 5: Comparison Results of Gender DIF using jMetrik, DIFAS and STATA**

| **Gender** | **Number of DIF** | | |
|---|---|---|---|
| | **jMetrik** | **DIFAS** | **STATA** |
| **Male** | 11 | 23 | 21 |
| **Female** | 30 | 24 | 21 |
| **Total** | **41** | **47** | **42** |

| **Effect Size** | **Number of DIF** | | |
|---|---|---|---|
| | **jMetrik** | **DIFAS** | **STATA** |
| **A** | 17 | 17 | 12 |
| **B** | 12 | 8 | 9 |
| **C** | 12 | 22 | 21 |

| | | | |
|---|---|---|---|
| **Total** | **41** | **47** | **42** |

**Table 6: Comparison Results of DIF Effect size using jMetrik, DIFAS nad STATA**

It can be seen from Table 5 that, the same results were obtained by each of the jMetrik and STATA programs, whiles DIFAS identified 47out of the 50 items as DIF. jMetrik identified more DIF items in favour of females (30 items) than males (11 items) whiles STATA identified an equal number (21 items) of DIF items for both male and female students.

Table 6 display the results of the effect size of items identified as DIF for the three software programs for evaluating DIF employing the Mantel-Haenszel method. From Table 5, it can be seen that DIFAS identified the high number of DIF items followed by STATA and then jMetrik. This implies that DIFAS is most sensitive when it comes to the identification of DIF. It is not surprising that DIFAS was also the software that exhibits the most C-level of DIF.

As regards output, jMetrik, STATA and DIFAS show a single table including all the results for all the items. One of the most important advantages of STATA and jMetrik over the other DIFAS program is the possibility of obtaining a graphical representation of the results, as well as being able to view them instantly as the analysis progresses. This means that a pragmatic researcher with little training would be able to interpret the results easily.

In conclusion, the comparison of the results from the three programs showed similarities in results were obtained using a thin matching strategy. In the case of STATA, its output did not include the effect size statistics. Items 3, 20, 40, 42 and 44 were not flagged by STATA and jMetrik software, while other items, such as 41, 43 and 45 (free of DIF), were flagged by DIFAS. Items 8, 17, 33 and 36 were also flagged by STATA and items 2,14, and 23 were flagged jMetrik. Finally, the different results provided by DIFAS, jMetrik and STATA can be attributed to how the

program implemented the thin matching strategy. The features of each software program have been shown to help the researcher choose depending on their interests.

## REFERENCES

Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research: Volume 1-The analysis of case-control studies*. Lyon: International Agency for Research on Cancer.

Camilli G., & Shepherd L. A. (1994). *Methods for identifying biased test items*. Sage, Thousand Oaks, CA.

Chalmers, R. P. (2018). Improving the Crossing-SIBTEST statistic for detecting non-uniform DIF. *Psychometrika, 83*(2), 376–386.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. *Differential Item Functioning*, *1*, 137-166.

González, A., Padilla, J. L., Hidalgo, M. D., Gómez-Benito, J., & Benítez, I. (2011). EASY-DIF: Software for analyzing differential item functioning using the Mantel-Haenszel and standardization procedures. *Applied Psychological Measurement*, *35*(6), 483.

Guilera, G., Gómez-Benito, J., & Hidalgo, M. D. (2009). Scientific production on the Mantel-Haenszel procedure as a way of detecting DIF. *Psicothema*, *21*(3), 492-498.

Hidalgo, M. D., Gómez-Benito, J. (2010). Education measurement: Differential item functioning. In Peterson, P., Baker, E., McGaw, B. (Eds.), *International encyclopedia of education* (3rd ed., pp. 36-44). Oxford, England: Elsevier Science & Technology Books.

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture. *Educational Psychology, 36*, 378–390.

Meyer, J. P. (2014). *Applied measurement with jMetrik*. London: Routledge.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297-334.

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti Estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement*, *40*, 353-370.

Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement, 29*(2), 150–151.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*(1), 23-37.

Sabatini, J., Bruce, K., Steinberg, J., & Weeks, J. (2015). SARA reading components tests, rise forms: technical adequacy and test design. *ETS Research Report Series, 2*, 1–20.

StataCorp, L. P. (1985). *Stata user's guide*. College Station, TX: Stata Press, Stata-Corp LP.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–364). Hillsdale, NJ: Lawrence Erlbaum.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, *26*(1), 55-66.