# Differential Item Functioning in Grade 8 Math Using Logistic Regression, Mantel-Haenszel and Logical Data Analysis

Jonas P. Villas
Mathematics Unit, Leyte Normal University, Tacloban City, Philippines

**Abstract**
This study identified biased test items through DIF analysis using Logistic Regression and Mantel-Haenszel Statistic. Biases were confirmed using LDA based on FGD's and interviews with teachers and students. The study made use of test scores from 99 male and 101 female grade 8 students to which 108 students were classified as low ability and 92 as high ability students based on their current English grades. A researcher-constructed and validated Statistics Achievement Test was used as research instrument based on Grades 7 and 8 Philippine K-12 competencies. The results from the two methods were compared, and it was found that sex and English ability bias exist. LDA reveals that bias favors females and high ability group in English which is associated with their capability of memorization and retention of topics taught procedurally. Recommendations include (1) incorporation of DIF analysis in test development; (2) the use of at least two methods in item bias detection; (3) the conduct of LDA or the qualitative component of DIF analysis is vital in understanding DIF to account for context specificity; and (4) for future research, the need to incorporate classroom observation as basis for LDA in DIF justification.

**Keywords:** item bias, item bias methods, logical data analysis, confirming bias, DIF
**DOI**: 10.7176/JEP/10-12-17
**Publication date**: April 30th 2019

## 1. Introduction

Test preparation means writing it in accordance to the rules of item construction, selecting items to be included according to TOS, reviewing the items, providing directions in answering the test, and deciding on the method of scoring. Further, analyses or revisiting the test to reassure that it measures what it purports to measure is necessary for a unified validity (Messick, 1994). However, the identification of difficulty, discrimination and reliability index is not sufficient to ensure test fairness (Gatchalian & Lantajo, 2010). The standards for educational and psychological tests require more analysis in the form of Differential Item Functioning (DIF) to identify further which item are appropriate to be administered to the examinees. *"DIF* refers to differences in item functioning after groups have been matched with respect to ability or attribute that the item purportedly measures" (Dorans & Holland, 1993). Bias happens when the results of the tests yield to "different meanings for scores earned by members of different subgroups." In order to detect these biases, a Differential Item Functioning (DIF) analysis has to be conducted.

Further, as we embark on One ASEAN Community, it is necessary to alleviate the standards and quality of achievement tests in the Basic Education platform. This is to ensure that we measure students' performance using tools appropriate to them and unstained with biases. With this in mind, there is a necessity to raise the consciousness of assessment practitioners to the unacceptability of tests to be administered having not undertaken bias studies. Item bias studies in the Philippines is at its "dearth," (Pedrajita & Talisayon, 2009).

These concepts, particularly testing DIF in a Mathematics Achievement Test, verifying whether bias occurs to subgroups in a population of students such as sex and English proficiency and identifying further why these items are flagged as bias, confronts the researcher. Moreover, knowing in refulgence the very cause of the existence of bias, external as it is - undoubtedly is imperative in the Philippine Educational System. In its process of educational reforms through K-12, this will provide a better understanding and improve the quality of teaching through fair assessment procedures.

The primary purpose of this study is to look into biased test items between male and female, and high and low English proficiency examinees in a researcher-constructed achievement test on Probability and Statistics strand in the K-12 program through DIF analysis. Further, as DIF simply flagged an item as bias, confirming such bias through a qualitative approach, e.g., Logical Data Analysis, is necessary for consideration in item revision and even in the entire pedagogical practice. Specifically, this study answers the following questions:

1. Is there a significant relationship between group membership and test performance on one item after controlling for total test scores?
2. Is there a difference in the *log odds ratio* of students in the focal and reference group?
3. What are the views of teachers and students on why certain items are flagged as biased to a particular group?

## 2. Assessment of Learning

"Assessment in education must first and foremost serve the purpose of supporting learning" (Black & William, 2006). Assessment is characterized as formative and summative. In the book of John Gardner (2006), formative assessment involves new ways to enhance feedback between students and teachers, and it requires new pedagogy and significant changes in classroom practice. Hence, what makes for effective learning means students have to be actively involved and that the results of an assessment are used in adjusting teaching and learning. On the other hand, a summative assessment is evaluative. Below are some tips in developing classroom assessment:

(1) Begin with the end in mind;
(2) Do not teach to the test but rather to the assessed content;
(3) Use a combination of summative and formative assessments.

This implies that a teacher must have a deep understanding of developing assessment tools as it requires innovation. Teacher's competencies in the Philippines include components on developing assessments (National Competency-Based Teacher Standards).

On the other hand, below are the "Standards for Teacher Competence in Educational Assessment of Students (Magno, 2013):

1. Teachers should be skilled in choosing assessment methods appropriate for instructional decisions;
2. Teachers should be skilled in developing assessment methods appropriate for instructional decisions;
3. The teacher should be skilled in administering, scoring and interpreting the results of both externally produced and teacher-produced assessment methods;
4. Teachers should be skilled in using assessment results when making decisions about individual students, planning to teach, developing curriculum, and school improvement;
5. Teachers should be skilled in developing valid pupil grading procedures, which use pupil assessments;
6. Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators; and
7. Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

Of these competencies, no. 2 and seven are what seemed to be set aside. Magno (2013) expounds that the bulk of assessment information teachers used for decision-making "comes from approaches they create and implement" and it "goes beyond readily available instruments." On the other hand, "fairness, the rights of all concerned, and professional ethical behavior must undergird all student assessment activities." This implies that teachers have to tailor cut their assessment tools to their type of students and be cautious about test validity. It is noted that if inappropriateness occurs in tests not limited to bias, the use of such should be put to rest.

Establishing test validity entails a rigorous process such as item analysis, the test of validity and reliability. Devine & Yaghlian of Cornell University defines item analysis as a systematic evaluation of the effectiveness of each item of a test and can tell us of the difficulty and discriminating power of an item and the effectiveness of each alternative. Further, they emphasized that for a test to be valid it should measure what it intends to measure. In establishing test validity, the test has to be subjected to content, criterion-related and construct validity. Moreover, a test is said to be reliable when it measures consistently what they were designed to measure; such methods would either be Test-retest method, Equivalent - forms method, Test-retest with equivalent forms, and Internal consistency method.

If a test satisfies item analysis on validity and reliability, does it immediately follow that the test is fair? If a teacher is mandated in NCBTS to exemplify the skill in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information, is it ethical, legal and appropriate that a test administered to students which are thought to provide the basis of grading or improvement in instruction are stained with bias? The identification of difficulty, discrimination and reliability index is not sufficient to ensure test fairness (Gatchalian & Lantajo, 2010).

## 3. Test Bias

The absence of fairness or bias happens when the psychometric properties of a test or when the manner in which it is used results different meaning for scores earned by members of different subgroups (Gatchalian, & Lantajo, 2010) and the existence of bias is an invalidity in itself or systematic error (Camilli & Shepard, 1994).

### *3.1 Differential Item Functioning*

To detect these biases a Differential Item Functioning (DIF) analysis has to be conducted. "DIF is an indicator that an item is potentially bias which will eventually provide a better understanding of the behavior of subgroups on potentially biased test items (Osterlind 1983) as cited by Gatchalian & Lantajo (2010). As an example, one could determine whether a particular test has a bias in gender, ethnicity or geographical location.

The study of Weiman (2001) reveals that "female excel in calculations, untimed and written tests, tend to have better grades in schooling, and until high school, females often exceed male on tests of math." This is contradicted in a way by Halpern (2000) as cited by Weiman (2001), “that in schools, males tend to excel at problem-solving, multiple choice tests and they outperform females in Standardized Achievement Tests." Looking back Hayd, Fennema & Lamon (1990) indicates that females’ outperformed males by only a negligible amount in mathematics performance, in 2010, Hayd, Lenberg &Peterson in their study concluded that males and females perform similarly in mathematics. The Big question in DIF is that, will this (e.g., gender) entail bias in an exam favoring to "Venus or Mars"? As DIF detects biases in an item in favor of a particular group, the question on whether male and female perform better in math is not a context within its scope. What it looks into are the biases in a test item and not into whether boys outperform girls or vice versa. A DIF study reveals that sex bias occurs in some items in reading comprehension (Salubayba 2012), in Philippine Aptitude Classification Test (Gatchalian, Lantajo, 2010), in Chemistry Achievement Test (Pedrajita and Talisayon, 2009), in Math and Science (Gierl, Khaliq and Boughton 1999). On another context, Gierl (1999) reveals that majority of multiple-choice sections in the Alberta Education Social Studies, 30 Diploma Examination composed of 70 items do not display DIF between male and female examinees.

### *3.2. DIF Detection*

The attributions to the existence of DIF do not rely only on the type of subject or on the type of tests in general but rather on how the items were constructed. Clauser & Mazor, (1998) as cited by Gierl, Khaliq & Boughton (1999), note that “DIF may be rooted to item impact or item bias. Item impact can be described as any group disparity in item performance that reflects actual knowledge and experience differences on the construct of interest” Alternatively, “item bias is defined as invalidity or systematic error in how a test item measures a construct for the members of a particular group” Camilli & Shepard, (1994). Hence, despite the many methods of detecting DIF, there is a necessity for these existences to be confirmed. Camilli & Shepard, (1994) suggested that DIF items can be confirmed bias through a qualitative measure of Logical Data Analysis (LDA).

DIF could be detected using several statistical tools and approaches. The characteristics of the dependent variables, e.g., dichotomous or polytomous are essential in determining DIF approaches, and the methods could be characterized as those relying on Item Response Theory (IRT) Model or Non-IRT Model. Most traditional methods used belong to the class of non-IRT methods which are designed to detect uniform DIF. Specifically, the MH, standardization, and SIBTEST procedures are based on statistics for contingency tables (Magis, Beland, Tuerlinckx and Boeck 2010).

*Logistic regression* (LR) can be seen as a method that bridges IRT and non-IRT (Camilli & Shepard, 1994). In logistic regression group membership, matching criterion and interaction effect between group and matching criterion are considered as matching variables in the mode. Logistic regression does not just detect bias but also determines uniform and non-uniform DIF. The process implored when one is using SPSS could be summarized as follows:

(1) Consider only the matching criterion in the model; then
(2) The main effect of the matching criteria and the group variable are considered, and model is reanalyzed; and
(3) The interaction between the matching and grouping criteria is analyzed in the third model.

These models are interpreted with respect to their chi-square value to determine uniform and non-uniform DIF. If one is interested in looking into the existence of DIF alone one could consider the *exp b* of the third model as this value is similar to the MH statistic and hence its odds ratio provided that its chi-square value is significant, e.g., less than 0.05. Taking the natural logarithm of this value and multiplying it to -2.35 will give us MH Delta which could be a basis of identifying the magnitude of DIF, Kamata, A & Vaughn, B. (2004).

"The MH method (Mantel & Haenszel, 1959) is prevalent in the DIF framework (Holland & Thayer, 1988). It aims at testing whether there is an association between group membership and item response, conditionally upon the total test score (or sum score)" as cited by Magis, Beland, Tuerlinckx, and Boeck (2010). The natural logarithm of the log odds ratio of MH statistics is also computed and multiplied to 2.35 to determine Delta MH. DIF magnitude is computed based on DIF classification rules which say that,

A – trivial or no DIF when the Delta MH is between 0 and 1
B – moderate effect, when Delta MH is equal or less than 1.5 but > 1
C – large DIF, if Delta MH is > 1.5. Kamata, A & Vaughn, B. (2004).

In a comparative study of Pedrajita and Talisayon, (2009) Logistic Regression (LR) and Mantel Haenszel (MH) are "widely implemented in detecting DIF," as this "procedures demonstrated the external validity." They also recommended that "LR and MH statistic be used in detecting DIF and that matching is conditioned simultaneously on total score, a categorical variable, and additional educational background variables like age, verbal ability, mathematical ability, social class, educational attainment, type of community and the like."

### *3.3 Logical Data Analysis*

Salubayba (2013) argued that though “empirical evidence of differential test performance is necessary” it is “not sufficient to enable any researcher to conclude about the presence of bias" and "the condition that the item is biased requires LDA." Salubayba (2013) added a qualitative dimension in detecting DIF by subjecting the examinees and the people involved with them in an interview and Focus Group Discussions (FGD).

These concepts, particularly testing DIF in a Mathematics Achievement Test, verifying whether bias occurs to subgroups in a population of students such as sex and English proficiency, and identifying further why these items are flagged as bias confronts the researcher. Moreover, knowing in refulgence the very cause of the existence of bias is undoubtedly imperative for the Philippine Educational System, in its process of educational reforms through K-12, and to provide better understanding and improve quality of teaching.

## 4. Conceptual Framework

Many studies on detecting biases across language groups have been conducted in the past, and with the K-12 using English as a medium of instruction in high school particularly in mathematics, it is daunting to ask whether proficiency in English have significant effects in the Mathematics performance of students.

In this study, math achievement test was administered to the matched groups of examinees (Figure 1). The scores were subjected to each of the DIF methods identifying bias in *gender* and *English proficiency.* The items flagged as bias were subjected to distracter analysis and its difficulty index was established. These, along with the DIF results were presented to the teachers and students to confirm bias and determine possible causes of such through Logical Data Analysis.

The following Null hypotheses were formulated:

a) There is no significant relationship between group membership and test performance on one item after controlling for total test scores.

b) There is no significant difference in the log odds ratio of students in the focal and reference group.

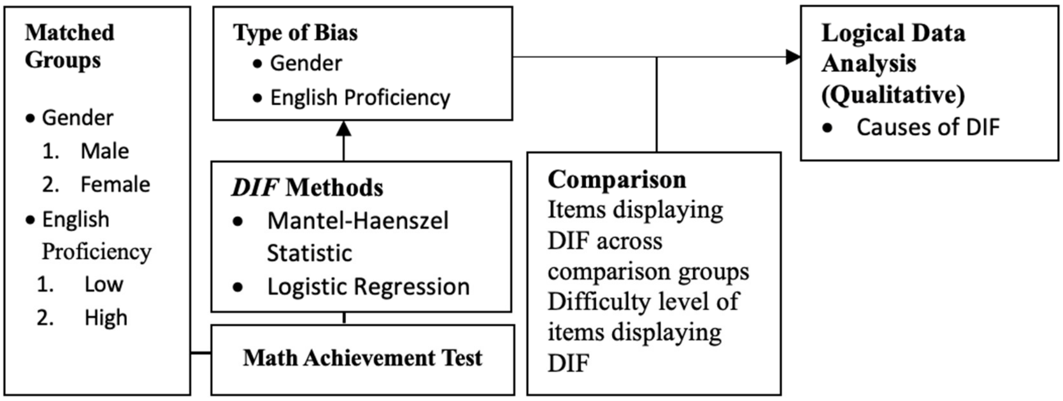

Figure 1. Methodological Flowchart of the Study

## 5. Methodology

The study provides a comparison between Mantel Haenszel Statistic and Logistic Regression Analysis in detecting DIF in an achievement test in probability and statistics among Grade 8 students in a particular high school in Metro Manila. This focuses on the performance of the aforementioned students grouped according to sex and English performance.

The study also implored a Logical Data Analysis to confirm and determine causes of DIF using Key Informant Interview and Focus Group Discussions among teachers and students guided by distracter analysis and the difficulty index of the DIF flagged items. In the conduct of Logical Data Analysis, the DIF items were reviewed to shed light in the reasons of bias existence. The conduct of FGD and a series of interviews was sequenced from teacher to student. The student FGD and interview findings were presented again to the teachers to provide possible causes of argument.

### *5.1 Sample and Sampling Procedure*

From a total of 560 Grade 8 students, the researcher sampled 200 students using a simple random sampling technique. These 200 students took the achievement test in probability and statistics.

The participants of the FGD include all grade seven and eight teachers, their math coordinator, and math expert, thus a total of 9 FGD participants. On the part of the students, a total of 12 participants representing, low, mid, and high scorers in the administered exam as well as male, female and high and low performance in English were considered in the determination of FGD participants.

### *5.2 Instrument*

The achievement test is a researcher made 40-item multiple choice examination. It was based on the K-12 competencies in the Probability and Statistics strand for Grades 7 and 8. A draft of the achievement test was constructed along with its table of specifications. It was presented to the teachers handling the subject in consultation with trainers in K-12 Mathematics component to establish content validity. Comprehensibility of the achievement test was assessed by the teachers as well. Also, since the study also implores qualitative measures in the form of Logical Data Analysis, Key Informant Interview and FGD guides were constructed to assist the researcher in LDA.

### *5.3 Data Collection Procedure*

The data collection procedure in the study was in three phases: (1) Consultation in the Development of the Achievement Test; (2) Administering the Examination; and (3) Logical Data Analysis.

The data collection commenced after approval for the conduct of the study was issued by the school administrator of where the study was conducted. Consultation and Orientation to teachers were conducted regarding the importance and the process adjured in the study. Both teachers and students were assured of the confidentiality of the data. Then after, presentation of the competencies in the Probability and Statistics Strand in the K-12 curriculum, as well as the draft examination, was conducted. A checklist was provided to verify relevance between the examination and the competencies. On the same draft, competencies and checklist were forwarded to the trainers of the K-12 Math Curriculum to establish content validity.

List of students in grade 8 was acquired, and the participants were determined using Simple Random Sampling. Students grades in the first to third grading period in the English subject were gathered and averaged to determine their English performance characterized as low and high performing students in English. The focal and referenced groups were then determined, examination scheduled and administered.

Data gathered was coded and analyzed using SPSS. Findings were then presented to the teachers and students to confirm and determine cause/s of bias through Key Informant Interview for Teachers and FGD for students. Findings from students and teachers were corroborated.

### *5.4 Data Analysis Procedure*

Statistical Package for Social Sciences was used in analyzing the data. Specifically, Mantel-Haenszel statistic, a non-parametric contingency procedure commonly used in detecting uniform DIF and Logistic Regression, which is used for a dichotomous scored variable, was entreated.

## 6. Results and Discussion

### *6.1 DIF Detection between Male and Female Examinees*

Table 1 shows the summary of flagged DIF items between male and female examinees and concept/skills measured in their content.

The grouping variable that was used is sex. MH Statistic reveals that 25 out of 40 multiple choice items display differential item functioning, that is $MH\chi^2$ yields a p-value less than 0.05 or 0.01, thus flagged as DIF and characterized as *large* evident by Delta MH greater than 1.5. An *odds ratio* greater than one across all flagged DIF is also observed implying that there is a direct or positive relationship between matched variable. Finally, the higher scores for the female examinees in the items signifies that DIF is in their favor. With this premise, the hypothesis of no significant relationship between group membership (sex) and test performance on one item after controlling for total test scores is rejected in favor of the alternative.

Table 1. *Summary of Flagged DIF items Between Male and Female Examinees and Concepts/Skills Measured in their Content*

| Item | Concept/Skill | $MH\chi^2$ | | LR | |
|---|---|---|---|---|---|
| | | Biased Against | Classification of DIF | Biased Against | Classification of DIF |
| 5 | Understanding and choosing the appropriate types of Graph given a data set. | Male | C | Male | C |
| 7 | Defining what a mean is. | Male | ***C*** | Male | ***B*** |
| 8 | Defining what a mode is. | Male | ***C*** | Male | ***B*** |
| 10 | Identifying midpoints of a class interval | Male | ***C*** | Male | ***A*** |
| 11 | Solving for the median given an ungrouped data | Male | C | Male | C |
| 12 | Solving for the weighted mean | Male | C | Male | C |
| 13 | Analyzing the mean and raw data | Male | ***C*** | Male | ***A*** |
| 14 | Solving for the median given an ungrouped data | Male | C | *Did not flag as DIF* | |
| 15 | Identifying the mode of ungrouped data | Male | C | Male | C |
| 17 | Solving for measures of central tendency given an ungrouped data | Male | ***C*** | Male | ***A*** |
| 18 | Analyzing the mean and raw data | Male | ***C*** | Male | ***A*** |
| 20 | Solving for the mode of a grouped data | Male | C | Male | C |
| 21 | Identifying the class size of a grouped data | Male | C | Male | C |
| 22 | Solving for the mean of grouped data | Male | ***C*** | Male | ***A*** |
| 23 | Identifying cumulative frequencies of a given interval | ***Male*** | C | ***Female*** | C |
| 24 | Solving for the median of a grouped data | ***Male*** | ***C*** | ***Female*** | ***A*** |
| 25 | Identifying total frequencies in a grouped data | Male | ***C*** | Male | ***B*** |
| 26 | Solving for the mode of a grouped data | Male | C | Male | C |
| 27 | Identifying the mode of ungrouped data | Male | C | Male | C |
| 28 | Solving for the median given an ungrouped data | Male | C | Male | C |
| 29 | Identifying the measures of variability | Male | ***C*** | Male | ***A*** |
| 30 | Understanding the definition of the measures of variability | Male | ***C*** | Male | ***B*** |
| 32 | Understanding the characteristics of the measures of variability | ***Male*** | C | ***Female*** | C |
| 34 | Solving for the range | ***Male*** | C | ***Female*** | ***A*** |
| 38 | Expressing frequencies to a percentage | Male | C | Male | C |

LR reveals 24 out of 40 multiple choice items have a p-value of less than 0.05, thus flagged as DIF. A DIF classification shows all 12 out 24 items displayed large DIF (C items), evident by delta greater than 1.5; 4 out of 24 items displayed moderate effect (B items) with delta greater than 1 but less than 1.5; and 8 out of 24 items displayed negligible DIF (A items) with delta less than 1. Of these items, 19 out of 24 are against male students and thus favor female while the remainder, e.g., 5, are against female and thus favor male with a negative or indirect relationship in the matched group. Specifically, these items are 23, 24, 29, 32, 34. The same items favored their counterpart when the data was analyzed using MH statistic. The hypothesis of no significant difference in the *log odds ratio* of students in the focal and reference group (male and female) is rejected in favor of the alternative.

### *6.2 DIF Detection between Low and High English Proficient Examinees*

Table 2 shows the summary of flagged DIF items between high and low English proficient examinees and concept/skills measured in their content.

Table 2. *Summary of Flagged DIF items Between High and Low English Proficient Examinees and Concepts/Skills Measured in their Content*

| Items | Concept/Skills | $MH\chi^2$ | | LR | |
|---|---|---|---|---|---|
| | | Biased Against | Classification of DIF | Biased Against | Classification of DIF |
| 2 | Identifying types of variable | *Did not flag as DIF* | | ***Low*** | ***C*** |
| 5 | Understanding and choosing the appropriate types of Graph given a data set | Low | C | Low | C |
| 7 | Defining what a mean is | Low | ***C*** | Low | ***B*** |
| 8 | Defining what a mode is | ***Low*** | ***C*** | ***High*** | ***B*** |
| 10 | Identifying midpoints of a class interval | ***Low*** | ***C*** | ***High*** | ***A*** |
| 11 | Solving for the median given an ungrouped data. | Low | C | Low | C |
| 12 | Solving for the weighted mean | Low | C | Low | C |
| 13 | Analyzing the mean and raw data | ***Low*** | ***C*** | ***High*** | ***A*** |
| 14 | Solving for the median given an ungrouped data | *Did not flag as DIF* | | ***High*** | ***C*** |
| 15 | Identifying the mode of ungrouped data | ***Low*** | C | ***High*** | C |
| 16 | Understanding the concept of a median | ***Low*** | ***C*** | *Did not flag as DIF* | |
| 17 | Solving for measures of central tendency given an ungrouped data | ***Low*** | ***C*** | ***High*** | ***A*** |
| 18 | Analyzing the mean and raw data | Low | ***C*** | Low | ***A*** |
| 20 | Solving for the mode of a grouped data | Low | C | Low | C |
| 21 | Identifying the class size of a grouped data | Low | C | Low | C |
| 22 | Solving for the mean of grouped data | ***Low*** | ***C*** | ***High*** | ***A*** |
| 23 | Identifying cumulative frequencies of a given interval | ***Low*** | C | ***High*** | C |
| 24 | Solving for the median of a grouped data | ***Low*** | ***C*** | ***High*** | ***A*** |
| 25 | Identifying total frequencies in a grouped data | Low | ***C*** | Low | ***B*** |
| 26 | Solving for the mode of a grouped data | Low | C | Low | C |
| 27 | Identifying the mode of ungrouped data | ***Low*** | C | ***High*** | C |
| 28 | Solving for the median given an ungrouped data | ***Low*** | C | ***High*** | C |
| 29 | Identifying the measures of variability | Low | ***C*** | Low | ***A*** |
| 30 | Understanding the definition of the measures of variability | Low | ***C*** | Low | ***B*** |
| 32 | Understanding the characteristics of the measures of variability | Low | C | Low | C |
| 34 | Solving for the range | Low | ***C*** | Low | ***A*** |
| 38 | Expressing frequencies to a percentage | ***Low*** | C | ***High*** | C |

MH statistics reveal that there were 25 out of 40 multiple choice items that have a significant $MH\chi^2$ value, e.g., it yields a p-value of less than 0.05, thus flagged as DIF and characterized as *large* evident by delta-MH greater than 1.5. An *odds ratio* greater than one across all flagged DIF is also observed implying that there is a direct or positive relationship between matched variable. Finally, high English proficient examinees scored higher in the items signifying that DIF is in their favor. The hypothesis of no significant relationship between group membership (English proficiency) and test performance on one item after controlling for total test scores is rejected in favor of the alternative. A DIF classification shows that all 25 items displayed large DIF against low English proficiency students. These items are: *5, 7, 8, 10, 11, 12, 13, 15, 16, 17, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 32, 34, and 38.*

Table 2 also shows the results of *logistic regression* which reveals that 26 out of 40 multiple choice items yields a p-value of less than 0.05, thus flagged as DIF. A DIF classification shows that 12 out 26 items displayed large DIF (C items), evident by delta greater than 1.5; four out of 26 items displayed moderate effect (B items) with delta greater than 1 but less than 1.5; and 10 out of 26 items displayed negligible DIF (A items) with delta less than 1. Of these items, 14 out of 26 are against low English proficient examinees, thus, favor high English proficient examinees while the remainder, e.g., 12, is against high English proficient examinees and, thus, favors low English proficient examinees with a negative or indirect relationship in the matched group. Specifically, these items are 8, 10, 13, 14, 15, 17, 22, 23, 24, 27, 28, and 38. The hypothesis of no significant difference in the *log odds ratio* of students in the focal and reference group (low and high English proficient examinees) is rejected in favor of the alternative.

### *6.3 Comparison of Methods*

Table 1 shows consistency between the Mantel Haenszel and Logistic regression in flagging items 5, 11, 12, 15, 20, 21, 26, 27, 28 and 38 as bias against male examinees with a magnitude of C. This implies that these items favor female. Items 7, 8, 25, and 30 were flagged as DIF by both Mantel Haenszel and Logistic regression in favor of female examinees. The magnitude of DIF as classified turns out to be different. The previous classify it as C or large DIF while the latter classify it as with moderate effect DIF or B. There were also discrepancies observed between mantel Haenszel statistic and logistic regression in its capacity of flagging an item as DIF. In this study three cases were observed:

(1) *Flagging DIF with different magnitudes.* Specifically, items 10, 13, 17, 18, 22, and 29 were flagged as DIF by MH statistic with a magnitude of C but flagged as DIF by LR with a magnitude of A. This DIF favors consistently to the same group, female.
(2) *Inconsistent DIF Flagging.* Item 14 was flagged as DIF by MH statistic but not by LR.
(3) *Flagging DIF in different favorability.* Items 23, 24, 32, and 34 were flagged as DIF differently by MH and LR statistic. It is different in the sense that all items flagged as DIF by MH statistics favor male, which contrary to what LR statistics reveal and with varying magnitude.

Similarly, Table 2 also shows consistency between the Mantel Haenszel and LR in flagging the items 5, 11, 12, 21, 26, and 32 as bias against low English proficient examinees with a magnitude of C. There were also items such as item 7, 18, 25, 29 and 34 that were flagged as DIF by both Mantel Haenszel and Logistic regression in favor of high proficient examinees in English but of varying magnitude. Specifically, MH and LR classified DIF for items 7 and 25 with magnitude C and B respectively, while MH and LR classified items 18, 28, and 34 with a magnitude of C and A respectively. Discrepancies were also noted, and this includes Items 2 and 14 flagged as DIF by LR statistic but not by MH. Similarly, item 16 was flagged as DIF by MH but not by LR favoring High English proficient examines classified as C. Items 8, 10, 13, 17, 22, 23, 24, 27 and 28 were flagged as DIF items by both MH and LR but with contradicting favored group. As MH results reveal that DIF favors high English proficient examinees while LR favors low proficient examinees.

Though Pedrajita and Talisayon (2009) suggest that MH and LR methods are consistent in identifying DIF, the above results take light in Karami (2011), wherein some techniques were not comparable under different sample size ratios and impact conditions in terms of Type I error and power. It was emphasized that at times, applying different DIF techniques will identify different items as displaying DIF.

### *6.4 Logical Data Analysis*

Whenever an item is flagged as displaying DIF, it is necessary to investigate if indeed the item is biased or not. Karami (2011) argued that “any item flagged as showing DIF is biased if, and only if, the source of variance is irrelevant to the construct being measured by the test," i.e., DIF is due to construct-irrelevant variance. As espoused by Messick (1994) the case of construct-irrelevant variance is anchored on the fact wherein “the groups of test takers perform differentially on an item, not because of an actual ability difference, but because of the unwanted effect of say a grouping factor."

In the logical data analysis, only items number **5**, 7, 8, 10, 11, 12, 13, 14, 15, 17, 18, **19**, 20, 21, 22, 25, 26, 27, **28**, 29, 30, and 38 was presented with potential bias based on the consistency in both methods and anchored on Camilli and Shepard’s (1994) notion of considerable judgment required about the relevance of DIF to the intended constructs. Difficulty index, when computed, reveals that only items 5, 19, and 28 are with optimum difficulty indices that need to be deleted from the roster of items. Student participants in the FGD were asked the question *of what comes into their mind when they were answering the item.* Their general answer was:

> *Hindi ko memorize [I was not able to memorize it]... Nakalimutan ko ang formula [I forgot the formula] ...*
> *Nakaligtaan ko ang proseso [I lost track of the process].*

On another hand, teachers provided with data results of distracter analysis and difficulty indices, merely confirmed the aforementioned items as bias. When asked why? They expressed that maybe the students forget the formulas. Also, when asked why a particular item favors female and those with high English proficiency, they said, because this group is good in memorization. During the FGD conducted there were no questions centered on how these concepts were taught. Nonetheless, it may be surmised that if these items were taught anchored on a model that requires memorization of definitions and algorithms, then these items may be indeed biased. However, if these items were taught innovatively using other models anchored on conceptual understanding and not on procedural, which requires memorization, then these items are not biased to the innate skills that female and high English proficiency students possess.

## 7. Conclusion and Recommendation

The results of DIF analysis showed that some items are statistically bias between male and female examinees as well as low and high English proficient test takers. Further, there seems to be an inconsistency in MH and LR in flagging bias items. On the other hand, items identified as bias were viewed by the students and teachers to be attributed to memorization. It was expounded that such skill is innate to those who are highly proficient in English and are mostly females. Concomitant to this, the following recommendations were made: (1) The practice of DIF analysis should be incorporated in test development to ensure test validity in a unified perspective; (2) At least two methods should be used in item bias detection to provide more justification on why the item is displaying DIF; (3) The conduct of logical data analysis or the qualitative component of DIF analysis is vital in understanding DIF to account for context specificity; and (4) For future research, there is a need to incorporate classroom observation as a basis for logical data analyses in DIF justification.

## References

Camilli, G. & Shepard, L. (1994). Methods for identifying biased items. Vol. 4, Sage Publication Inc.

Devine, M. & Yaghlian, N. (N.D.). Test construction manual: construction of objective tests. *Center for Teaching Excellence*. Retrieved from www.cte.cornell.edu

Dorans, N. & Holland P. (1993). DIF detection and description. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

Gatchalian, C. & Lantano, A. (2010). Revisiting the Philippine aptitude classification test: analysis of potentially biased items [online].

Gierl, M. (1999). Differential item functioning on the Alberta education social studies 30 diploma examination. *Canadian Social Studies*, Vol. 33, No. 2.

Gierl, M., Khaliq, S. & Boughton, K. (1999). *Gender differential item functioning in mathematics and science: prevalence and policy implications.* [online]. University of Alberta: Centre for research in applied measurement and evaluation.

Holland, P., & Thayer, D. (1988). Differential Item Performance and the Mantel-Haenszel Procedure. In H. Wainer, and H. I. Brown (Eds.), Test Validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hyde, J. S., Fennema, E., & Lamon, S. (1990). Gender differences in mathematics performance: A meta-analysis. Psychological Bulletin, 107, 139 –155.

Kamata, A. & Vaughn B. (2004). An introduction to differential item functioning analysis. *Learning Disabilities: A Contemporary Journal* 2 (2), 49-69.

Karami, H., & Nodoushan, M. (2011). Differential item functioning (DIF): Current problems and future directions. *International Journal of Language Studies*, 5(3), 2011 (pp. 133-142).

Linberg, S., Hyde, J., & Petersen J. (2010). New trends in gender and mathematics performance: A meta-analysis. Psychology Bulletin. 136(6): 1123–1135. doi: 10.1037/a0021276

Magis, D., Béland, S., Tuerlinckx, F., & Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. Behavior Research Methods 2010, 42 (3), 847-862 doi:10.3758/BRM.42.3.847.

Magno, C. (2013, July). Standards of teacher competence on student assessment in the Philippines. [online] The Assessment Handbook, 10(42). Manila, Philippines: De La Salle University

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.

Messick (1994). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Research Report* RR-94-45. Educational Testing Service, Princeton, N.J.

Pedrajita, J. &Talisayon, V. (2009). Identifying biased test items by differential item functioning analysis using contingency table approaches: a comparative study. [online]*Education Quarterly*, 67(1). UP College of Education.

Salubayba, T. (2013). Differential item functioning detection in reading comprehension test using

Mantel-Haenszel, item response theory, and logical data analysis. [online] *The International Journal of Social Sciences*, 14 (1). Retrieved from www.Tijoss.com.

Weiman, H. (2001). Gender differences in cognitive functioning. Retrieved from http://faculty.kendall.edu/hweiman/GenderDifferences.html.