DIFFERENTIAL ITEM FUNCTIONING DETECTION IN READING COMPREHENSION TEST USING MANTEL-HAENSZEL, ITEM RESPONSE THEORY, AND LOGICAL DATA ANALYSIS

Teodora M. Salubayba
PhD in Research and Evaluation, University of the Philippines, Diliman
Director of the Institutional Statistics and Evaluation Center
Behavioral Statistics Professor
St. Scholastica’s College, 1004 Manila, Philippines

Abstract

The Mantel-Haenszel (MH) and Item Response Theory One-Parameter Logistic (IRT-1PL) model were investigated for their sensitivity in detecting (differential item functioning) DIF in Reading Comprehension test administered to 1925 Grade Six pupils. The pupils were from six mixed-gender and three all-girl schools run by religious congregation in the Philippines. The two DIF procedures revealed that the focal group (boys in gender-based DIF, girls in mixed-gender schools in school type-based DIF) was disadvantaged in most of the items. Large and consistent differences between the performances on the test items of the focal group and the reference group were evident under IRT-1PL as also depicted in the item characteristic curves (ICC). MH detected fewer DIF items that resulted to more DIF-free items. On the contrary, IRT-1PL detected more DIF items that produced brief test instrument. Items flagged as DIF were subjected to qualitative logical data analysis (LDA), a review process done by the test users, practitioners, and the examinees themselves. LDA involved focus group discussions (FGD) and interviews. Responses from the FGD and interviews confirmed biased items and revealed causes of DIF items like words and phrases in the items, options that favored one group, familiarity with the reading selections, among other things. Recommendations include reviewing the reading selections in the test, item writing procedures, considering response patterns and choice of distracter in DIF procedures, and creating better test evaluation practices to ensure fair tests and assessments.

Keywords: Differential Item Functioning, Item Response Theory, Item Characteristic Curves, Mantel-Haenszel, Qualitative Logical Data Analysis

Introduction

The process in developing test instruments that is ‘fair’ for all test takers requires the removal or revision of items that are displaying or showing differential item functioning (DIF). In an achievement test, items are said to be showing DIF when equally able examinees in terms of ability and drawn from the same population but belonging to different subgroups such as male or female, do not have the same chances of correctly responding to an item. Test developers should strive therefore to identify and remove symbols, words, phrases, and contents that are generally regarded as causing DIF particularly in reading comprehension test as far as certain groups of a given population are concerned.

There is a need for a more sophisticated way of evaluating test instruments particularly those that attempt to detect DIF items. In the study, the Mantel-Haenszel (MH) chi-square under the classical test theory (CTT) and the Item Response Theory (IRT) one-parameter logistic model were considered as DIF procedures for critical evaluation of test items, for detecting the presence of DIF in the items. These two DIF procedures found to be effective and sensitive in detecting DIF in the items. Several studies (Navas-Ara & Gomez, 2002; Nijenhuis, 2004; Sheppard t al, 2006; Stoneberg, 2004; Wolfe & Phyllis, 1990; et al) made use of Mantel-Haenszel in detecting DIF. The
procedure has been found simple as it does not require highly specialized software. On the other hand, IRT-1PL procedures have been found to be more sensitive in detecting DIF and can produce precise, valid, and relatively shorter instrument (Baker, 2004; Edelen & Reeve, 2007; Embretson, 2000; Hambleton, 1991; et al.). In such cases, the choice of better DIF procedure is left to the informed judgment of the test users and policy makers.

Qualitative Logical Data Analysis

Items flagged as DIF are considered potentially biased items against either the focal group or the reference group. A DIF item can be confirmed bias after the qualitative logical data analysis because according to Camili & Sheppard (1994), empirical evidence of differential test performance is necessary, but not sufficient to enable any researcher to draw conclusion about the presence of bias; the condition that the item is biased requires a logical data analysis. Logical data analysis (LDA) is meant to discover detectable patterns of DIF or common characteristics of individual items and the causes of DIF in the items. Causes of DIF in the items according to Reynolds (2006) could be the familiarity of one group with the topic or clues in the items compared to other group, and student’s experience in terms of school curriculum that contributed to test item performance.

In this study, LDA was performed after obtaining significant DIF statistics to conclude that items really are biased and to determine the causes or origins of DIF for the subgroups under study. Causes of DIF in the item were discovered and emerged through interviews and discussion with the examinees and the people involved with them. The use of LDA had enriched the analysis of DIF and added a qualitative dimension to the study, in contrast to the highly quantitative procedures that are normally involved in detecting the presence of DIF in reading comprehension test.

DIF Analysis Comparison Groups in the Study

DIF procedures involve subgroups referring to the focal group and the reference group. Several studies that made use of DIF procedures considered gender – male vs. female (Sheppard, et al, 2006; Stoneberg, 2004), racial vs. ethnic (Meiring, 2006; Nijenhuis, 2004), and public vs. private schools (Pedrajita, 2007); but not single-sex schooling which is one of the subgroups of interest in the present study in comparison to mixed-gender schooling in an achievement test. In addition to gender (between girls and boys), this study also intended to investigate how DIF affects the test performances of girls in all-girl schools and girls in mixed-gender schools in reading comprehension.

DIF analysis was not only applied in achievement tests, aptitude tests, and ability tests, but also in survey questions about disabilities (Koretz, 2008), in personality inventory (Sheppard et al, 2006), in test of citizenship skills (Baghi & Ferrara, 1990). The DIF analysis in the present study examined the items in English Reading Comprehension Test. The said test was used as a vehicle in comparing the performances of Mantel-Haenszel and IRT One-parameter logistic model.

The results of the quantitative and qualitative DIF procedures in this study aim to represent a substantial new body of knowledge to help explain the causes of the gender-based and school type-based DIF in reading comprehension test. The findings of this study are hoped to contribute to research and practice in schools’ and institution’s testing program, the formulation and implementation of educational policies and decisions related to test development, and help test developers and test users to make informed decisions regarding the selection of test item evaluation procedures, particularly in the area of differential item functioning.

Method

Research Design

This study provided a comparative analysis of the two DIF procedures, the Mantel-Haenszel and IRT-1PL in their effectiveness and sensitivity to detect DIF in test items in the context of the English Reading Comprehension test administered to Grade six pupils. Besides the two aforementioned highly technical statistical procedures, the study also employed a qualitative logical data analysis to confirm biased items and explore the causes of DIF. The comparison focused on the performance on the test items of the focal group (boys in in gender-based
DIF; girls in the mixed-gender schools in school type-based DIF) and the reference group (girls in gender-based DIF; girls in the all-girl schools in school type-based DIF).

Research Participants and Source of Data

The study utilized the results of the reading comprehension test administered to 1925 Grade 6 pupils in School Year 2010-2011 in three all-girl schools and six mixed-gender private schools run by a religious congregation in the Philippines. The schools were similar in terms of certain characteristics such as school orientation, mission-vision, class size, pupil-teacher ratio, professional development opportunities for teachers, adequacy of facilities, and school curriculum. The 1925 Grade Six pupils consisted of 1395 girls (72%) and 530 boys (28%). Out of 1395 girls, 571 (41%) were from the all-girl schools and 824 (59%) from the mixed-gender schools. Of the 1354 students in the mixed-gender schools, 824 (61%) were girls and 530 (39%) were boys.

Assessments and Measures

The Reading Comprehension test was an existing instrument administered yearly since 1997 to approximately 2000 Grade 6 pupils in the nine schools. It has undergone pilot testing but no serious study was carried out to establish its psychometric properties. The 60-item test consisted of dichotomously scored multiple-choice items and contained thirteen passages or reading selections that were typical of the kind of materials found in the grade six textbooks. The reliability scale analysis of the test yielded Cronbach’s coefficient alpha of .878 which indicates that 87.8% of the observed variance is true variance. The quantitative factor analysis and qualitative review of test items by English teachers resulted to two subskills namely, literal recognition (40 items) and inferential comprehension (20 items). Literal recognition items measured pupil’s ability to locate and identify ideas or information explicitly stated in the reading selection. The answers to the items measuring literal recognition were not always the exact words or phrases in the reading selection but some were paraphrased in the answer choices.

Inferential comprehension items measured pupil’s ability to use ideas and information explicitly stated in the reading selection; his intuition and experiences in arriving at a conclusions and making conjectures about the passage.

DIF Analyses

The Mantel-Haenszel (MH) was one of the two DIF procedures used in this study to identify the presence of DIF in a given item. It used the MH-chi-square statistics and MHD-values to determine the presence of DIF in the test item and the group that is disadvantaged on the item. In the study, the computed MH chi-square value was compared to the critical value using alpha of .05 to determine its significance. A significant M-H statistics (p<.05) suggested the presence of DIF in the item.

Item Response Theory One-parameter Logistic Model (IRT 1PL) was one of the IRT models used to test the presence of DIF in the item. The R software was used in determining item parameters, group variance-covariance, and also provided the item characteristic curves (ICC) for the comparison groups. The formula $\chi^2 = b^2_{diff}/v(b_1) + v(b_2)$ (Hambleton, 1991) was used to compute for the IRT chi-square. An item was flagged as DIF when the $\chi^2$ value is significant at .05 (p<.05).

The Focus Group Discussion (FGD) Guide, Interview Guide, and the DIF Checklist were administered in preparation for the LDA that was conducted to determine why certain items seemed to be relatively more difficult to a particular group, the focal group under gender-based and school type-based DIF analysis.

The FGD asked questions relating to the test performance of pupils in reading comprehension. It employed a number of different interview approaches based on the progress of the focus group discussion but included unstructured, hierarchically focused, narrative and critical incident approaches. An open-ended guide was used to facilitate the discussion.

The group of psychometricians and test practitioners were the item reviewers. They accomplished a checklist containing the list of DIF items in the test. They also deliberated and concluded that the DIF items are really biased against a
particular group, and determined the possible causes of DIF in the items, such that group membership influenced pupils’ performance on the test.

Figure 1 illustrates a procedural flowchart that guided the study in carrying out the DIF procedures and in determining the causes of DIF in the test items. MH and IRT-1PL were statistical DIF procedures, while LDA employed qualitative approaches. The separate results of MH and IRT-1PL in terms of DIF-loaded items were compared. The resulting DIF items were subjected to logical data analysis meant to discover the causes of DIF and confirmed biased items.

Moreover, findings show that the focal group considered as the primary interest in DIF analysis was disadvantaged compared to the reference group in the study. In Table 2, there are more items that potentially biased the focal group than the reference group.

Table 2 shows that in almost all the analysis, out of the total DIF items, there are more items that disadvantage the focal group than the reference group under IRT-1PL in gender-based and school-type based DIF. Under MH, the proportion of DIF items that disadvantage the focal and the reference group are almost the same.

IRT-1PL DIF procedure provides the item characteristic curves (ICC) of the focal and reference group that illustrate the disparity between the performances on the items of these comparison
groups. Results of the gender-based DIF showing the DIF items with greater $\chi^2$ value are presented in Figure 2.

Figure 2 shows that in the first five DIF items based on the size of $\chi^2$ values, the reference group leads on the items; the ICCs for the reference group lie above the ICCs for the focal group.

Similar findings are found between school types (all-girl and mixed-gender schools) as depicted in Figure 3. The ICCs of the reference group lie above the ICCs of the focal group.

Furthermore, agreements and disagreements between the results of MH and IRT-1PL were sought. There were a number of DIF items that were detected and common under MH and IRT across comparison groups in gender-based and school type-based DIF. By gender, there were five common DIF items and by school type, there were 14 common DIF items. However, disagreements were observed between the results of IRT and MH. There were a large number of DIF items in IRT but were considered DIF-free in MH, and vice versa as presented in Tables 3 and 4.

Table 3

<table>
<thead>
<tr>
<th>Intersecting DIF Items in IRT and MH</th>
<th>Disjoint Set of DIF Items Detected in IRT only</th>
<th>Disjoint Set of DIF Items Detected in MH only</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1, 2, 4, 9, 11, 15, 19, 22, 25, 26, 28, 32, 38, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 56, 57, 59, 60, 55, 52, 61, 63</td>
<td>2, 3, 6, 11, 12, 14, 15, 16, 18, 19, 21, 22, 23, 24, 25, 26, 27, 33, 34, 36, 37, 38, 39, 40, 43, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 56, 57, 59, 60, 30, 35</td>
</tr>
</tbody>
</table>

Total = 5

Total = 41

Total = 4

Table 4

<table>
<thead>
<tr>
<th>Intersecting DIF Items in IRT and MH</th>
<th>Disjoint Set of DIF Items Detected in IRT only</th>
<th>Disjoint Set of DIF Items Detected in MH only</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1, 2, 4, 9, 11, 15, 19, 22, 25, 26, 28, 32, 38, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 56, 57, 59, 60</td>
<td>6, 8, 12, 18, 20, 23, 24, 27, 29, 33, 34, 36, 37, 38, 41, 42, 43, 44, 45, 46, 51, 53, 54, 60</td>
</tr>
</tbody>
</table>

Total = 14

Total = 22

Total = 11

In gender-based DIF, items identified under the subskills of reading comprehension showed that there were more DIF items under the literal recognition that disadvantaged the focal group than the reference group. Similarly, in inferential comprehension, almost all the DIF items were to the
disadvantage of the focal group and these items were referring to long reading selections. With regard to school type, almost all the DIF items in literal recognition and in inferential comprehension were to the disadvantage of the focal group; only a single item put the reference group at a disadvantage.

Table 5
*Set of DIF Items in Reading Comprehension Subskills under MH and IRT by Gender*

<table>
<thead>
<tr>
<th>Reading Comprehension by Gender</th>
<th>LR (40 Items)</th>
<th>IC (20 Items)</th>
<th>LR (40 Items)</th>
<th>IC (20 Items)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IRT</td>
<td>IRT</td>
<td>MH</td>
<td>MH</td>
</tr>
<tr>
<td>Potentially Biased Against the Focal Group (Boys)</td>
<td>31</td>
<td>13</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Potentially Biased Against the Reference Group (Girls)</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Total Item with DIF</td>
<td>32 (80%)</td>
<td>14 (70%)</td>
<td>4 (10%)</td>
<td>5 (25%)</td>
</tr>
<tr>
<td>Total Item without DIF</td>
<td>8 (20%)</td>
<td>6 (30%)</td>
<td>36 (90%)</td>
<td>15 (75%)</td>
</tr>
</tbody>
</table>

Note: LR – Literal Recognition  
IC – Inferential Comprehension

Table 6
*Set of DIF Items in Reading Comprehension Subskills under MH and IRT by School Type*

<table>
<thead>
<tr>
<th>Reading Comprehension by School Type</th>
<th>LR (40 Items)</th>
<th>IC (20 Items)</th>
<th>LR (40 Items)</th>
<th>IC (20 Items)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IRT</td>
<td>IRT</td>
<td>MH</td>
<td>MH</td>
</tr>
<tr>
<td>Potentially Biased Against the Focal Group (Mixed-gender)</td>
<td>22</td>
<td>13</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Potentially Biased Against the Reference Group (All-girl)</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Total Item with DIF</td>
<td>23 (58%)</td>
<td>13 (65%)</td>
<td>17 (43%)</td>
<td>8 (40%)</td>
</tr>
<tr>
<td>Total Item without DIF</td>
<td>17 (42%)</td>
<td>7 (35%)</td>
<td>23 (58%)</td>
<td>12 (60%)</td>
</tr>
</tbody>
</table>

Note: LR – Literal Recognition  
IC – Inferential Comprehension

Characteristics of the DIF items were also explored with regard to their difficulty level. Table 7 shows that not only difficult items were prone to DIF but items with moderate and easy difficulty level as well.

Table 7
*Distribution of DIF Items According to Difficulty Level Gender-Based and School Type-Based DIF in Reading Comprehension*

<table>
<thead>
<tr>
<th>DIFFICULTY LEVEL</th>
<th>Gender-Based DIF</th>
<th>School Type-Based DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Literal Recognition</td>
<td>Inferential Comp.</td>
</tr>
<tr>
<td></td>
<td>IRT</td>
<td>MH</td>
</tr>
<tr>
<td>Very Difficult/Very Hard</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Difficult/ Hard</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Desirable/Medium</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>Easy</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Very Easy</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TOTAL DIF Items</td>
<td>32</td>
<td>4</td>
</tr>
</tbody>
</table>
Moreover, correlation between the difficulty level of the DIF items and DIF values was also examined. In gender-based DIF, MH results showed significant correlations between the inferential comprehension items’ difficulty level and the corresponding DIF values; a moderate positive correlation coefficient \( r = .598 \) was obtained. In IRT, there was moderate negative correlation \( r = -.588 \) between the difficulty level of the inferential comprehension items and the DIF values. The correlation coefficients were significant at alpha level of .05. Negative and positive correlations as found in IRT and MH meant that the easier the item, the higher the DIF values. The findings implicate that the easier the inferential comprehension item the more likely the presence of DIF in the item. No significant correlation was found in other analysis, in school type-based DIF.

The possible causes of DIF were determined through qualitative logical data analysis (LDA). The DIF items were reviewed by the test practitioners and psychometricians who were directly involved with the Grade six pupils. Results of the LDA shed light on why the items were loaded with DIF. LDA was also done with regard to the responses of the pupils in focus group discussion (FGD) about DIF items.

For instance, logical data analysis showed that the distinctive cause of gender-based and school type-based DIF in the test was paraphrasing the correct answer in the literal recognition items. The item reviewers also mentioned that some of the answer choices were problematic, thus option analysis was recommended as among the possible areas for future study.

The observations of the item reviewers and the responses of the pupils during the FGD found that most of the reading selections catered to the interest of the reference group more than the focal group. In gender-based DIF, the pupils from the focal group mentioned that most of the reading passages were more “boring” and they took them long, thus, the length of the reading selection influenced the performance of the focal group. The pupils in the focal group also mentioned that the time allotted to answer all the items was not enough. They also found some of the questions confusing. In items that measured literal recognition, the focal group exhibited weakness in paraphrased correct option compared to the reference group. With regard to school type-based DIF, in literal recognition and inferential comprehension, the focal group was found to be at a disadvantage in answering the items compared to the reference group.

Overall, the causes of DIF in the reading comprehension items were: (1) paraphrasing the correct options from the reading selection that the pupils especially the focal group seemed to be weak at, (2) unfamiliar words and phrases in the stem of the item that seemed to be more apparent among the members of the focal group both by gender and school type, (3) unfamiliar words and phrases in the reading selection that hampered the pupils to perform better in the test especially the members of the focal group across comparison groups, (4) familiarity with the scenario of the reading selections that favored the reference group, and (5) the length of the reading selection that put the focal group at a disadvantage in gender-based DIF because the boys seemed to be impatient in reading longer passages.

**Discussions**

The results of the present study that IRT identified more items displaying DIF than MH may be attributed to the nature and assumptions of the two DIF procedures besides what Edelen & Reeve (2007) have underscored regarding IRT that produce brief instrument. IRT and MH have different assumptions and variance of measurement errors (Grujtejer and Kamp, 2008). They also have different focus. MH focuses on the test and is sample dependent (Hambleton, 1991). IRT focuses on the item by modeling the response of an examinee of given ability to each item in the test (Baker, 2004; Demars, 2010; Embretson, 2000). Among other things, IRT
and CTT are different in score interpretations and measure of ability (Embretson, 2000). However, these two procedures have the advantages compared to the other DIF procedures as reported in the studies. The Mantel Haenszel has the greatest advantages as a DIF detection technique because it provides the best results (MJ Navas-Ara, 2002), an adequate substitute for the one-parameter IRT if the sample size is at least 750 (Baghi & Ferrara, 1990), it is conceptually uncomplicated, and it doesn’t require highly specialized software (Camili & Shepard, 1994), and found to have its strengths in detecting DIF over other methods (Nijenhuis, 2004; Stoneberg, 2004). On the other hand, IRT-DIF approach was shown to be flexible and powerful method for differential item functioning (DIF) detection but computationally-complicated, requiring many model-refittings. (http://thepsychometricscentre.com/dif-detection)

Comparisons between the use of MH and IRT DIF procedures in detecting DIF have been examined (Embretson, 2000; Dorans & Kulick in Wolfe & Phyllis, 1990; Baghi & Ferrara, 1990; MJ Navas-Ara; 2002; Koretz, 2008) in several contexts and in different academic subjects and areas.

IRT-1PL and MH produced different test lengths. MH that detected less DIF items resulted to longer test with more DIF-free or acceptable items. IRT that detected more DIF items generated brief test instrument with a few DIF-free or acceptable items. The test practitioners therefore may decide on what to consider given the results of the two DIF procedures.

Performances of the comparison groups are also different under the two DIF procedures. In MH there is almost equal set of items favoring either the focal group or the reference group. IRT-1PL results are more forthright showing that one group is disadvantaged on the item. There are marked differences between the items disadvantaging either group. Results show that there were passages regarded as potentially biased for the focal group in gender-based DIF like “two girls sharing a room to study and trying to reach a compromise” and “a passage on good storytelling.” In these passages, all the items were found as potentially biased against the focal group. Moreover, the reference group seemed to relate easily to the item that asked to classify the books about “Make Your Own Greeting Cards, Prayers, and Pet Care”. Classifying the aforementioned books may be easier for the pupils in the reference group and they appear to have more library skills than focal group. On the other hand, the passage “Spiders and Moon” was found to be easier for the focal group to capture the ideas and visualize the story. The same holds true in school type-based DIF, the focal group is disadvantaged on the items compared to the reference group. The IRT-1PL therefore is more capable of detecting DIF items that distinctively favoring one group than the other.

With regard to subskills in the test, it is apparent that literal recognition items are more susceptible to DIF compared to inferential comprehension. As mentioned literal recognition items measured pupil’s ability to locate and identify ideas or information explicitly stated in the reading selection. Inferential comprehension items measured pupil’s ability to use ideas and information explicitly stated in the reading selection; his intuition and experiences in arriving at a conclusions and making conjectures about the passage. The findings that literal recognition items are more prone to DIF may be attributed to the fact that stored knowledge may influence more the performance on the items. Thus, examinees of similar ability may perform differentially as influenced by group membership that inadvertently provides exposure to the information and ideas in the reading selections.

Agreements and disagreements between the results of the two DIF procedures are apparent in gender-based and school type-based DIF. There are a number of intersecting DIF items as well as disjoint set of DIF items in the two tests across comparison groups. It is therefore upon the informed judgment of the test practitioners to ascertain the properties of the items that potentially biased the comparison groups; to investigate the components of the item that require further content review.

Exploring the causes of DIF is an important task towards developing fair assessments; they are the key to decisions about item revisions and/or removal. Several causes of DIF in the test items as identified by the item reviewers revealed that focal group is weak in detecting the correct answer to literal recognition items because of paraphrased options in the answer choices. Difficulty in answering inferential comprehension items is influenced by unfamiliarity with the words and phrases in the stem of the item as well as drawing
conclusions from the reading selections. This is evident when the items are examined between the focal and reference group in gender-based and school type-based DIF analysis.

The comments of the item reviewers in the study show concordance that because of time allotted to finish the test, the pupils seemed to be anxious to answer all the items. They tend to make blind guess to the items referring to lengthy reading selections. This was apparent in gender-based DIF among the pupils in the focal group who preferred to read first the question before the passage, thus resulted in guessing the correct answer to the item. The causes of DIF therefore are the paraphrased correct options, words and phrases in the stem of the item, length of the reading selection, and the scenarios of some reading selections that the item reviewers considered as catering to the interest of the reference group. Furthermore, the attitude of the boys (focal group in gender-based DIF) toward reading long passages is also evident as causing DIF in the items.

Gender-based DIF can be brought about by gender stereotype and different exposures and experiences of the girls and boys. Wolfe & Phyllis (1990) have stated that test items may have contents that influence performance in the test such that when questions are set in experiences more familiar to one sex than the other, one may have the advantage in answering the item than the others given the same ability.

Societal expectations such as girls have to be exposed to particular interests and boys to the interests and habits presumably masculine may have impact on the language they learned and ideas they imbibed and internalized during their growing up years. Societal expectations are those expectations from family members especially parents, and the surrounding environment including school teachers. One way in which the social environment influences the learning of languages differently along gender lines are through adults’ discrimination between the sexes after the infant’s birth (Santrock, 1998 as mentioned in Madu, 2010).

Moreover, several studies have already mentioned gender differences in test performance. Li, (2010) has reported that females are more serious and careful in taking reading tests while the males are more impatient and easily get bored. Li also noted that the readers’ metacognitive awareness of reading strategies is closely linked to their language proficiency.

Items that put the girls in the all-girl school (reference group in the school type-based DIF) at a disadvantage were pointing to the same passage. This implicates that the girls were affected by the passages not necessarily the items. This may also mean that the passages may not be within the interests, experiences, and exposure of the girls; that they might have a different schema compared to the girls in the mixed-gender schools. Experiences and exposures of these two groups of girls may have influenced their responses on the reading comprehension test items. The school type, that is, being in the mixed-gender or in the all-girl school may be a factor that contributed to the differences in test performance between girls of the same reading comprehension ability. Studies shows that girls in single-sex schools are more exposed in reading, devote much time in reading (Nidoy, 2011), and learn better in single-sex environment (Herr & Arms, 2004; Hoffman & Badgett (2008); Riordan, 2007). Single-sex schools are seen as benefitting and advantaging girls academically (Spielhofer, 2004; Daly & Deffy, 2004; Flowers, 2005; Debare, 2004)

The reading selections or the passages considered in the test may have contexts that were related to the experiences and exposures of the girls in the school where they were in. With regard to the items, how the stem of the items were formulated, the options provided after each item, may have also contributed to the performance differences on the test of these girls. Since the girls in the mixed-gender school show a slight weakness in dealing with the items that were spread out from different passages, the nature of the weakness seemed to be on the comprehension of the question and choosing the correct option.

Studies have explicited how orientations, experiences, and exposures of a particular group can influence performance on the test; that may either put them at the advantage or at a disadvantage in responding to set of test items. These factors may have direct impact on test performance. Gender and school type as two grouping variables in the study were deemed to influence pupils’ performance on the test items.

In addition to the aforementioned differences in the performances across comparisons
groups and their implications to research in education, the item reviewers in the study also mentioned the manner by which some of the items were written, the stem and the answer choices that contributed to DIF in the items. These are the following:

a. The manner by which the multiple choice items were formulated may have violated some rules in writing multiple-choice items and options. The item reviewers noted some of the seemingly problematic options that contributed to DIF in the items like "neither a or b, either b or c, both c and d, and none of the above."

b. The fill-in-the-blank in the stem of the item is also a potential cause of DIF according to the item reviewers.

c. There are items that are prone to guessing because of problematic options according to the item reviewers. The pupils also expressed that a few items can be answered without reading the passage in the reading comprehension test.

Other significant findings of the study concern the difficulty level of the items. Items displaying DIF are representative of very hard to very easy items. There is almost equal number of DIF items from very easy to very hard under the literal recognition and inferential comprehension skills. This means that not only difficult or very difficult items are susceptible to DIF but the easy and very easy items as well.

Moreover, significant correlations found between DIF values in gender-based DIF and the level of difficulty of the inferential comprehension items implicates that information about the difficulty level of the items can be the basis of predicting DIF in the items. This conclusion is only true in inferential comprehension items between the two groups in gender-based DIF but not under school type-based DIF.

As this study makes clear the advantages and disadvantages of the two DIF procedures that resulted in very different conclusions regarding the DIF items in the tests, each can provide unique information about evaluating test items that could be useful in both research and testing practices decision making. Under ideal conditions, it would be desirable to consider both approaches when examining achievement test items.

This study is limited in scope like other studies. There are many different possible areas by which researchers and test practitioners may consider in detecting DIF in achievement test items. Recommendations for future research using the two DIF procedures include improving the process of test evaluation to ensure fair tests and assessments. Other grouping variables may be considered in future DIF analysis besides gender and school type like socioeconomic status, school location, and other possible grouping variables that are deemed to influence performance in the test holding ability constant.

Future studies may consider options in the analysis of DIF in the test items. Besides a correct answer, other concerns regarding the responses to a multiple-choice item include: (1) omitting the item, (2) not reaching the item, and (3) choosing a distractor. Response pattern of the pupils on each item can be considered in the investigation of DIF that may help define the nature of DIF associated with a particular item.

References


