Determining Differential Item Functioning in Mathematics Word Problems Using Item Response Theory

Teodora M. Salubayba
St. Scholastica's College-Manila
dory41@yahoo.com

Abstract

Mathematics word-problem test was evaluated for differential item functioning (DIF) using Item Response Theory One-Parameter Logistic model. The test measured skills in application of one to two steps basic math operations, computation skills involving fraction, percentage, ratio, proportion, and complex analysis involving two or more steps in complicated problem solving. It was administered to 1925 Grade Six pupils in six mixed-gender private schools and three all-girl schools in the Philippines. Results revealed that the focal group (girls in gender-based DIF, girls in mixed-gender schools in school-type-based DIF) was disadvantaged in most of the items. The item characteristic curves showed remarkable differences between the test performances of the comparison groups. Focus group discussions and interviews with the participants confirmed DIF items and revealed the causes of gender-based and school-type-based DIF. DIF in favor of each gender showed agreement with the sex-role stereotypes; DIF in favor of each school type conformed to the findings of earlier studies about differences in the experiences of the girls in the all-girl schools and their counterparts in the mixed-gender schools. Recommendations include reviewing test item writing procedures in mathematics word-problems, and creating better test evaluation practices to ensure fair tests and assessments.

Key Concepts: *Differential Item Functioning, Item Response Theory, Focal Group, Reference Group*

Introduction

Differential item functioning (DIF) analysis is one of the several processes that are used in test development to ensure that items are free from bias. It is conducted to investigate how items function in various subgroups (Schnipke, et al, 2000). It is viewed to occur when examinees of equal proficiencies, but coming from different populations differ in the probability of answering an item correctly (Roussos & Stout, 2000).

The main purpose of this study was to identify math word problems items affected by DIF when given to these subgroups, that is, boys vs. girls in mixed-gender schools under gender-based DIF; girls in all-girl school vs. girls in mixed-gender school under school type-based DIF. The Item Response Theory One-Parameter Logistic (IRT- 1PL) model was used to identify DIF in the test items. IRT DIF technique has been found sensitive in detecting DIF and can produce precise, valid, and relatively shorter instrument (Baker, 2004; Edelen & Reeve, 2007; Embretson, 2000;Hambleton, 1991; et al). Besides the statistical DIF technique, the causes of DIF in the items were also determined through focus group discussion and interviews with the examinees and those directly involved with the pupils.

The present study aims to contribute a substantial new body of knowledge to help explain the causes of the gender-based and school type-based DIF in math word problems; and to help test practitioners make informed decisions regarding test item evaluation procedures particularly in the area of differential item functioning.

.

DIF Analysis Comparison Groups

DIF analysis procedures involve subgroups referring to the focal group and the reference group. In the study, under gender-based DIF analysis, the focal group comprised the girls; the reference group – the boys. With regard to school-type-based DIF, the focal group consisted of the girls in the mixed-gender schools; the reference group pertains to the girls in the all-girl schools.

Method

Research Participants and Source of Data

The study utilized the results of the Math word problems test administered to 1925 Grade 6 pupils in three all-girl schools and six mixed-gender private schools. 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.

Instrument

The 40-item mathematics word-problems test measured subskills in (1) application of simple one to two basic math operations, (2) application of computation skills involving fraction, percentage, ratio, and proportion, and (3) application of complex analysis involving two or more steps in complicated problem solving.

DIF Analysis

Item Response Theory One-parameter Logistic Model (IRT-1PL) used the IRT-chi-square statistics which was computed after the item parameter estimates for each group were placed on a common scale. An item was flagged for DIF when the χ^2 value is significant at .05 (p<.05). After the DIF items had been identified in IRT, they were subjected to a qualitative data analysis to determine and explore the causes of differential item functioning in Math word problems. The set of DIF items were examined and compared qualitatively to confirm biased items, providing a spectrum of opinions regarding the causes of bias in the items. Causes of DIF in the item were discovered and emerged through interviews and focus group discussion (FGD) with the examinees and the people involved with them.

The use of qualitative data analysis had enriched the analysis of DIF and added a qualitative dimension to the study, in contrast to the highly quantitative techniques that are normally involved in detecting the presence of DIF in achievement tests. The confluence of the quantitative and qualitative DIF procedures is expected to provide a clear picture of the biased items in Math word-problems test including the causes of bias.

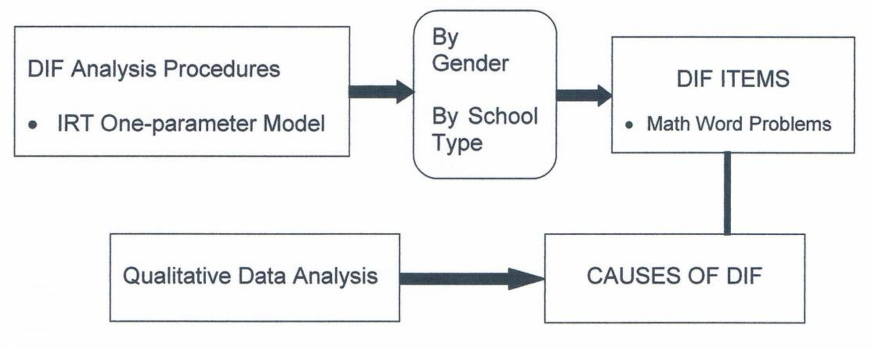

Figure 1. Conceptual Framework

Figure 1 shows the IRT-1PL DIF procedure used to detect DIF items in Math word problems. The arrow from the DIF analysis procedures points to the box on the DIF items in Math word problems. This study aimed to identify items that are loaded with DIF using the IRT-1PL. It was hypothesized that there would be differences in the test performance of girls and boys in the mixed-gender schools, girls in the all-girl schools and girls in the mixed-gender schools in the test items. It was believed that group membership influenced the pupils' performance on the test. This implies that there are certain factors such as culture and environment that influence the extent to which DIF occurs in subject-oriented tests and subtests.

Results

The IRT-1PL resulted to a number of gender-based and school type-based DIF items in Math word problems. In Table 1, findings reveal that there are more DIF items than DIF-free items under school type-based DIF. In gender-based DIF, there is almost the same number of DIF items and DIF-free items.

Table 1

Results of DIF Analysis across Subgroups

	Gender-Based DIF	**School Type-Based DIF**
DIF Items	19 (48%)	26 (65%)
DIF-free items	21 (52%)	14 (35%)
TOTAL	40	40

Table 2 shows that out of 26 DIF items under school-type-based DIF, the focal group considered as the primary interest in DIF analysis was disadvantaged compared to the reference group. However, the reverse is true under gender-based DIF; the reference group was disadvantaged in most of the items (68% out of the total 19 DIF items).

Table 2

Set of DIF Items that Potentially Biased the Focal Group/Reference Group

Comparison Group	**Gender-Based DIF**	**School Type-Based DIF**
Focal Group	6 (32%)	25 (96%)
Reference Group	13 (68%)	1 (4%)
Total DIF Items	19	26

Note:
Focal Group: Girls in Gender-based DIF; Girls in Mixed-Gender Schools in School Type-based DIF
Reference Group: Boys in Gender-Based DIF; Girls in All-Girl Schools in School Type-based DIF

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 the comparison groups. The DIF items were ranked based on computed IRT χ^2 values. Results of the gender-based DIF showing the three DIF items with greater χ^2 value are presented in Figures 2 and 3. Findings show that the reference group (boys) is disadvantaged on item 12; the ICC lie below the ICC of the focal group (girls). However, in the two items, item 8 and item 29, the reference group is at the advantage, the ICCs lie above the ICCs of the focal group. In gender-based DIF as shown in Figure 2, it appears that the focal group is disadvantage on hard and difficult items (item 8 & 29) that require higher computation skills whereas the reference group on easier item (item 12) that measures application of one to two simple basic math operations. Under school types as presented in Figure 3, all-girl vs mixed-gender schools as comparison groups, findings show that the reference group is advantage in the three items with high χ^2 values (items 6, 20, & 3); the ICCs of the reference group lie above the ICCs of the focal group.

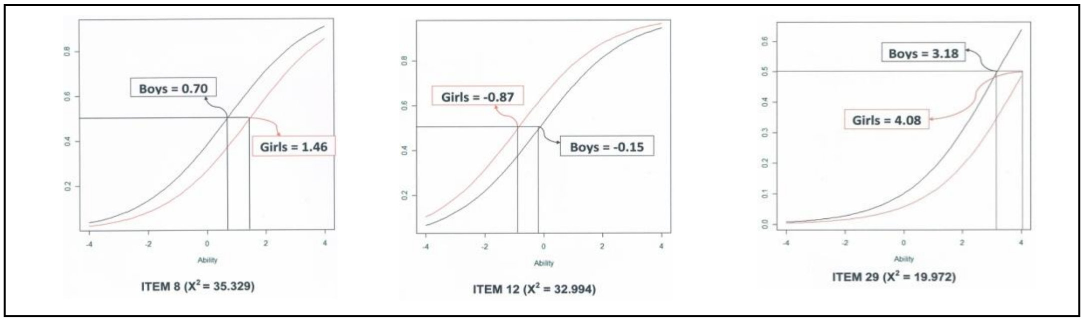

Figure 2. Item Characteristic Curves of the Three DIF Items by Gender for the **Focal Group** (Girls) and the **Reference Group** (Boys)

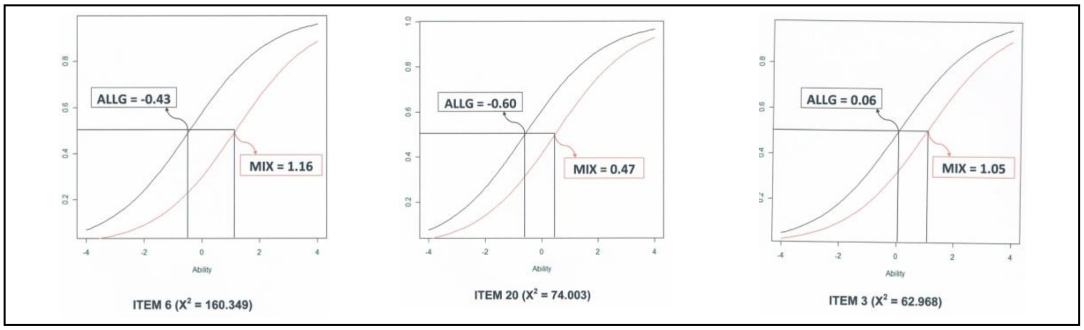

Figure 3. Item Characteristic Curves of the Three DIF Items by School Type for the **Focal Group** (MIX-Girls in Mixed-Gender Schools) and the **Reference Group (**ALLG**-**Girls in All-Girl Schools)

Furthermore, DIF items identified under the subskills of Math word problems showed the following results. In gender-based DIF, the reference group was at a disadvantage in items that measured the application of simple one to two steps basic math operations. The focal group was at a disadvantage in items measuring the application of computational skills involving fractions and percentages, also in items that measured the application of complex analysis skills involving two or more steps in complicated problem solving. Results are shown in Table 3.

Table 3

Set of DIF Items in Math Word Problems Subskills by Gender

Gender-based DIF	**Subskill 1 (14 Items)**	**Subskill 2 (12 Items)**	**Subskill 3 (14 Items)**
Potentially Biased Against the Focal Group(Girls)	1	2	3
Potentially Biased Against the Reference Group (Boys)	11	0	2
Total Item with DIF	12 (86%)	2 (17%)	5 (36%)
Total Item without DIF	2 (14%)	10 (83%)	9 (64%)

Note:
Subskill 1: F1SBO – Application of One to Two Steps Basic Math Operations
Subskill 2: F2AFPPR – Application of Computation Skills Involving Fraction, Percentage, Ratio, and Proportion
Subskill 3: F3ACAS – Application of Complex Analysis Skills Involving Two or More Steps in Problem Solving (Complicated Problem Solving)

In Math application by school type as presented in Table 4, it was consistent that the focal group was at a disadvantage in the three subskills compared to the reference group.

Table 4

Set of DIF Items in Math Word Problems Subskills by School Type

Math Application by School Type	**Subskill 1 (14 Items)**	**Subskill 2 (12 Items)**	**Subskill 3 (14 Items)**
	IRT	IRT	IRT
Potentially Biased Against the Focal (Mixed-gender)	4	11	10
Potentially Biased Against the Reference (All-girl)	1	0	0
Total Item with DIF	5 (36%)	11 (92%)	10 (71%)
Total Item without DIF	9 (64%)	1 (8%)	4 (29%)

Note:
Subskill 1: F1SBO – Application of One to Two Steps Basic Math Operations
Subskill 2: F2AFPPR – Application of Computation Skills Involving Fraction, Percentage, Ratio, and Proportion
Subskill 3: F3ACAS – Application of Complex Analysis Skills Involving Two or More Steps in Problem Solving (Complicated Problem Solving)

Characteristics of the DIF items were also examined with regard to their difficulty level. Findings show that not only difficult items were prone to DIF but items with moderate and easy difficulty level as well. This observation is apparent in all the subskills across comparison groups under gender-based and school type-based DIF.

The possible causes of DIF were determined through item reviews by the people knowledgeable in testing and directly involved with the Grade six pupils, the school psychometricians, guidance counselors, and selected Math teachers. Responses of the pupils in focus group discussion (FGD) about DIF items were also considered important in determining the causes of DIF. Results of the qualitative review of DIF items shed light on why the items were loaded with DIF.

Under school-type-based DIF, one of the causes of DIF in the math word problems pertains to the focal group lacking mastery of certain skills when the test was administered. These skills were computation analysis involving fractions and percentages as well as skills like performing two or more basic operations and identifying number sentences in the problems presented. In gender-based DIF, the length and the scenario of the math word problems had put the reference group at a disadvantage. Both in two DIF analyses, the causes of DIF items were: (1) terms in the problems presented that were confusing to the focal group, and (2) the time allotted to answer all the items that was not enough according to the focal group. The pupils in the reference group mentioned that they were trained to answer the test, an achievement test in particular, in the fast way they can. Other concerns that the item reviewers noted were the stem of the items and the answer choices that may have put a particular group at a disadvantage.

Discussion

Gender and school type as grouping variables have different impact on the test. There are marked differences between the test performances of the comparison groups. In school type-based DIF, the focal group is disadvantaged on most of the items compared to the reference group. Although, it is different in gender-based DIF where the reference group is disadvantaged on easier items, the results show that the focal group is disadvantage on the more difficult items involving higher computational skills.

Under gender-based DIF, the focal group (girls) compared to the reference group (boys) is found weak on the items that require higher computational skills involving fraction, percentages, and combinations of more than two mathematics operations. However, the reference group (boys) is disadvantage on items measuring one to two simple basic operations with contexts that are unfamiliar to them like measuring ingredients for baking and cooking. Wolfe & Phyllis (1990) state 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. Study shows that although the boys are performing better in Mathematics compared to girls (Ercikan & Barnett, 2006), girls may outperform the boys when the problems are context-specific. On the other hand, the reference group (boys) leads on items that entail higher computational skills and exhibits greater interest in complicated Math word problems, resulting to better performance on these items compared to the focal group.

Differences in the cognitive abilities of females and males are best described in the study of Halpern (2004). Kimball (1989) as mentioned in Halpern (2004) hypothesized that girls' learning is more rote than boys' learning, so girls' learning is assessed best with familiar problems.

The difference between the performances of the girls in two school types appear to be influenced by the Math remediation and Math enrichment provided to the pupils in the all-girl schools. The Math teachers in all-girl schools mentioned during the interviews that they tried to find time to provide remedial class to the low performing pupils. Math enrichment is provided for the high performing pupils. A few studies and articles about single-sex and mixed-gender classrooms (e.g. Nidoy, 2011) have shown different experiences and exposures of girls in the two classroom environments.

Moreover, the item reviewers in the study mentioned the manner by which some of the items are written, the stem and the answer choices that apparently contributed to DIF in the items. They noted some of the seemingly problematic options like “neither a or b, either b or c, both c and d, and none of the above.” The fifth option “NA” (or not available) found to be appealing to the pupils in the focal group. During the FGD with the pupils, it was mentioned that the NA option was their best option when they get tired of thinking and computing, and when they were in a hurry to finish the test. The fill-in-the-blank in the stem of the item is also a potential cause of DIF according to the item reviewers.

The difficulty level of the DIF items is also significant findings of the study. Items displaying DIF in the test 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 three subskills of Math word problems. This means that not only difficult or very difficult items are susceptible to DIF but the easy and very easy items as well.

Given that DIF is inadvertent and apparently inevitable despite carefully written items, people in testing and evaluation need to be more cautious and continue evaluating any test for DIF to ensure fair tests and assessments, and to arrive at a fair evaluation of individual examinee. Recommendations for future research using the DIF techniques include improving the process of test evaluation to ensure fair tests and assessments. Standard procedures in writing test items should be considered in reviewing further the items. The study suggests that researchers conduct similar studies using the IRT DIF techniques in other tests besides the Math word problems. 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 also consider options in the analysis of DIF in the test items. Considering other item parameters like the guessing parameter may be compared to the results of DIF.

References

Baker, F. & , Kim, S. H. (2004). Item Response Theory Parameters Estimation Techniques. New York: Marcel Dekker Inc. 2nd edition.

Camilli and Shepard (1994). Methods for Identifying Biased Test Items. London: SAGE Publications.

Demars, C. (2010). Item Response Theory. Ney York: Oxford University Press Inc.

Edelen, M. O. & Reeve, B. (2007). Applying Item Response Theory (IRT) Modeling to Questionnaire Development, Evaluation, and Refinement. *Quality Life Research*. 16: 5-18.

Ercikan, K. & Mendes-Barnett S. (2006). Examining Sources of Gender DIF in Mathematics Assessments Using a Confirmatory Multidimensional Model Approach. *Applied Measurement in Education*. 19(4), 289-304.

Embretson, S. E. & Reise, S. P. (2000). Item Response Theory for Psychologists. London: Lawrence Erlbaum Associates, Publishers.

Gruijter, D. M. & Kamp, L. J. (2008). Statistical Test Theory for the Behavioral Sciences. New York: Taylor & Francis Group.

Halpern, D. F. (2004). *A Cognitive-Process Taxonomy for Sex Differences in Cognitive Abilities.* American Psychological Society: Current Directions in Psychological Science. Retrieved July 1, 2011 from ProQuest Educational Journals

Hambleton, R.K., Swaminathan, H., & Rogers, J.H. (1991). Fundamentals of Item Response Theory (IRT). London: Sage Publications.

Nidoy, R. (2011). Advantages of Single-Sex Schooling: Explanation of Teachers Who Taught in Both Coed and Single-sex Schools. *Parents for Education Foundation. Paper presented during the Third International Congress on Single-Sex Schools*. Poland: European Association for Single-Sex Education (EASSE).

Roussos, L. A. & Stout, W. F. (2000). *A Formulation of the Mantel-Haenszel Differential Item Functioning Parameter with Practical Implications.* Retrieved March 2, 2010 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1a/70/b1.pdf

Schnipke, D. L., Roussos, L. A., Pashley, P. J. (2000*). A comparison of Mantel-Haenszel Differential Item Functioning Parameters.* Retrieved March 2, 2010 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1a/70/b4.pdf

Tyre, P. (2005). Boy Brains, Girl Brains. *Newsweek.* Vol. 146, Issue 12, p. 59. Retrieved September 12, 2006. EBSCO Host Research Databases.

Walker, C., Zhang, B., & Surber, J. (2008). Using A Multidimensional Differential Item Functioning Framework to Determine If Reading Ability Affects Student Performance in Mathematics. *Applied Measurement in Education.* Retrieved July, 2012 from ProQuest Journals.

Wolf, L. R. & Phyllis, R. (1990). The SAT Gender Gap. *Women and Languge*. Vol. 13, Issue 2. Retrieved September 26, 2011 from ProQuest Journal