Educational Measurement and Evaluation Review (2016), Vol. 7 Issue 2

Test Development Using Differential Item Functioning

Arlene N. Mendoza
Pangasinan State University, Binmaley

Elsie M. Pacho
Don Mariano Marcos Memorial State University, Bacnotan

Abstract
Differential item functioning (DIF) analysis is an essential element in the evaluation of the fairness and validity of educational tests. This study developed a researcher-made test utilizing four DIF detection models: Mantel-Haenszel Chi-Square Statistic, Logistic Regression, Transformed Item Difficulty, and Rasch Model. Descriptive-comparative research design was employed in the DIF analysis based on students' differences on age, sex, language ability, socio-economic status, and school type. The study made use of the test scores of 188 BSE students major in Mathematics in the validated Achievement Test in Calculus I which was used as research instrument. Results of the study revealed that the revision and elimination of the potentially biased items in the test resulted to a valid, reliable, and fair test.Further, Mantel-Haenszel was the least sensitive in detecting DIF items among the models utilized. Moreover, the IRT Models, particularly the Rasch Model, revealed the highest number of detected DIF items, hence, has the highest statistical power of detection in the test constructed.

Keywords: Item bias, Development of an Achievement, Test, Rasch Model, Transformed Item Difficulty,Logistic Regression, Mantel-Haenszel Chi-Square Statistic

Introduction

In any assessment situations, one of the major goals of test developers is to ensure that the test instrument is free from bias against any identifiable groups. Bias is a major factor for tests considered unfair, inconstant, and contaminated by extraneous factors. A test is biased against or for a particular group if it under-predicts or over-predicts, respectively, their performance on the criterion of interest relative to some other groups (Pedrajita & Talisayon, 2009). Educational or psychological tests may include items that operate differently for certain groups. It is important to identify these items because they may lead to unfair results for groups being compared. The reason for such items to operate differently may be gender, age, culture, school type, teaching practices, classroom size, socio-economic status, or language differences between groups.

There are several methods of evaluating item bias, including the use of sensitivity reviews, differential validity studies, and Differential Item Functioning (DIF) detection methods (Wood, 2011). This study focused on item bias detection in an Achievement Test using differential item functioning (DIF) detection methodsfor test improvement. Differential item functioning (DIF) analysis is typically used to identify test items that are differentially difficult for respondents who have the same level of knowledge, skill, or ability but differ in ways that should be irrelevant to their performance on the test. The presence of large numbers of items with DIF is a severe threat to the construct validity of tests and the conclusions based on test scores derived from items with and items without DIF (Karami, 2012). Various differential item functioning (DIF) procedures have been proposed to assess potential bias. Despite the widespread application of DIF analysis in psychometric circles,it seems that the inherent complexity of the concepts in DIF analysis has hampered its wider application among less mathematically oriented researchers and only a limited number of them appearto be in current use. Thus, this study attempted to utilize these methods in assessing a

dichotomously scored test to detect bias test items and consequently construct a reliable, valid, and fair test.

This studyaimed to develop an Achievement Test by detecting biased test items particularly in Calculus usingItem Response Theory-based (IRT) model via Rasch Model and Transformed Item Difficulty Approach. In addition, two types of Classical Test Theory Models or the Contingency Table Approach via Logistic Regression and Mantel-Haenszel Chi-Square Statistics were also employed. According to Bradley (2009, p.5), "IRT techniques are the 'gold standard' of DIF detection." However, in the study of Salubayba (2013), she found out that Mantel-Haenszel and IRT-1PL were found both effective and sensitive in detecting DIF in the items. She showed that grouping variables like gender and school type were deemed to influence the performance of the pupils in reading comprehension and math application.However, in the study conducted by Madu (2012) to assess gender-related DIF using Transformed Item Difficulty, results show an incorrect picture of the quality of education for different groups and this may likely lead to the resources for education being distributed in an unfair manner. On the other hand, Pedrajita and Talisayon (2009) found out that there was a high degree of correspondence between the Logistic Regression and the Mantel-Haenszel Statistic in identifying biased test items. These findings gave the researchers an idea on applying an IRT-based models and Contingency Table Approach in detecting potentially bias item.

Further, comparativeanalyses among these DIF methods were done based on their sensitivity of detecting biased items. Moreover, the effect of biased items' elimination on the construct, content, and concurrent validity, and internal consistency reliability of the achievement test were determined.This study was delimited to some contextual variables such as age, sex, language ability, socio-economic status, and school type. In most situations, these factors were observed to affect examinees' chance to succeed in each test item.

The Achievement Test constructed focused on the topics about Calculus in order to construct set of valid,

reliable, and unbiased test items that wouldalso help dealing with problems monitoring the dynamical changes of biological samples, all kind of optimization problems or economic problems.Besides the significant aspect that this part of mathematics helps in development of an analytical mathematical thinking, calculus proves its effectiveness by solving real, practical problems. Calculus is used to find the rate of change; hence, it is very important because our society relies on it.

This study can significantly contribute to educational research especially in test development. Test experts, developers, and educators may: (1) gain insights on the applicability of DIF detection methods; (2) realize the validity of DIF methods in detecting biased test items based on students' differences on their age, gender, language ability, socio-economic status and school type; (3) use DIF methods in developing valid and equitable tests; and (4) employ DIF methods in purifying their assessment instruments.

Method

Research Design

This study employed the descriptive-comparative research design utilizing a researcher-made Achievement Test in Calculus. Development of the test was done by detecting item bias using the four methods of differential item functioning (DIF) models: Mantel-Haenszel Chi-Square Statistic, Logistic Regression, Transformed Item Difficulty, and Rasch Model. The DIF analysis of the test items were based on the students' differences on age, sex, language ability, socio-economic status, and school type. The detected biased items in the test using the four methods were revised and some were eliminated as based on the criteria set by the model. The validity and reliability of the test were then computed afterwards. The statistical power of detection of the DIF models was also determined through their sensitivity in detecting DIF items based on thesegroup

differences. The more DIF items detected, the higher the statistical power of detection of the DIF Models.

Participants

The test was administered to 188 college students taking up Bachelor of Secondary Education major in Mathematics from different Higher Education Institutions (HEIs) in Region I, private and public, who have already taken up their Calculus course, and enrolled during the first semester of SY 2014-2015.

Materials/Instrument

A questionnaire was formulated which solicited information regarding the students' age, sex, grade point average in Calculus I and in English I, socio-economic status, and school type. This information served as basis in detecting biased items. In addition, a researcher-made achievement test in Calculus I was constructed which consisted of 100 items. This is a multiple-choice test which covered concepts on Functions (8 items), Limits and Continuity (27 items), Derivatives (31 items), and Analysis of Functions and their Graphs (34 items).

Procedure

The researcher constructed an achievement test in Calculus Iand was evaluated by a panel of experts in the field of Mathematics who are at least Master's degree holder in Mathematics and have been teaching Calculus for at least five years. After validation, the test was administered to a group of BSE students major in Mathematics for field testing. The pilot testing has been conducted to a group of Bachelor of Science in Education majoring in Mathematics in Higher Education Institutions (HEI's) which were not included in the study.

When the test was tested for validity and reliability, it was administered again to 188 college students taking up

Bachelor of Secondary Education majorin Mathematics from different HEI's in Region I.The students were randomly assigned as the focal group and the reference group. The matched groups were based on the type of school they came from (public or private), their sex (male or female), their age (17 and below or 18 and above), grade point average in English I (above or below average of the group performance) and their socio-economic status in terms of gross monthly income (Php 8,000.00 and below or above Php 8,000.00). These groups were used as the bases in detecting DIF items through the DIF methods. The detected DIF items were then revised and improved if not eliminated. The revised version of the test was again subjected to test validity and reliability.The comparisons among the DIF methods were also done afterwards.

Data Analysis

This study has employed two Item Response Theory (IRT) DIF detection methods, the Transformed Item Difficulty approach, and the Rasch Model. Likewise, two Classical Test Theory (CTT) approaches were also considered, Logistic Regression, and Mantel-Haenszel Chi-Square Statistics. The efficacy of the methods was compared based on their sensitivity on detecting DIF items.

In calculating the MH statistics, the first step is to compute the probabilities of correct and incorrect responses for both groups. The second step is to find out how much more likely are the members of either group to answer correctly rather than incorrectly to the item. The overall DIF is calculated by summing the odds ratios at all ability levels and dividing them by the number of ability levels. The resulting index is the Mantel-Haenszel odds ratio denoted by α_{MH}. This index is usually transformed by the following: $\beta_{MH} = \ln\alpha_{MH}$ (Karami, 2012). A negative β_{MH} indicates DIF in favor of the focal group whereas a positive MH Δ shows DIF favoring the reference group (Wiberg 2007). Sometimes, β_{MH} is further rescaled into:

$$MHD = -2.35 \ln\alpha_{MH}.$$

A positive MHD indicates that the item was more difficult for the reference groups and a negative value shows that the focal group faces more difficulty with the item (Karami, 2012).

In Logistic Regression, an item is classified as displaying DIF if the two-degree-of-freedom Chi-squared test is beyond 5.9915 tested at 0.05 alpha significance and has a p-value less than or equal to 0.01 (set at this level because of the multiple hypotheses tested). Moreover, the Zumbo-Thomas (ZT) effect size measure had to be at least an R-squared of 0.130 (Zumbo, 1999). For ZT effect size measure, items were categorized as "A" if the value of their R-squared is significantly different from 0 and less than 0.13. Also, items were categorized as "B" if R-squared differ from 0.13 and less than 0.26. And it is considered under category "C" if R-squared differ from 0.26 and less than 1.

In Transformed Item Difficulty Approach, items with a perpendicular distance $(|D_i|)$values in excess of 1.5 reveal DIF. The larger (D_i) is, the more biased the item. A signed transformed difficulty measure of DIF, which preserved both the direction and magnitude of DIF was obtained by attaching a positive sign to (D_i) if the item reveals DIF in favor of the focal group, and a negative sign if the item reveals DIF in favor of reference group. For this study, a value of (D_i)greater than 1.5 indicates DIF,favoringthe focal group, whereas a value (D_i) less than -1.5 indicates DIF favoring the reference group.

The detection of differential item functioning through Rasch Model was performed using the Lord's chi-square method with one parameter logistic model. In this study, only one parameter was used ; hence, the Lord's chi-square with one parameter logistic model permits us to get item parameter estimates from the Rasch or one-parameter logistic (1PL) model. The calculated chi-square statistic was compared to a critical value (3.8415) based on an *a priori* specified level of significance (0.05), with degrees of freedom (df) corresponding to the number of parameters examined

for each item. If the observed chi-square exceeds the critical value, then the null hypothesis of no DIF is rejected.

Table 1 summarizes the detection threshold and the effect size of each DIF models in detecting DIF items.

Table 1
Detection Threshold and Effect Size of the DIF Detection Methods

DIF Detection Methods	Detection Threshold	Effect Size	Code	Scale Used
Mantel-Haenszel Chi-Square Statistics	3.8415	0.0 – 1.0	A	Delta Scale
		1.0 – 1.5	B	
		> 1.5	C	
Logistic Regression	5.9915	0.0 – 0.13	A	Zumbo and Thomas (ZT)
		0.13 – 0.26	B	
		0.26 – 1.0	C	
		0.0 – 0.035	A	Jodoign and Gierl (JG)
		0.035 – 0.07	B	
		0.07 – 1.0	C	
Transformed Item Difficulty	>1.5 and < -1.5	MHD value	N/A	N/A
Rasch Model	3.8415	0.0 – 1.0	A	Delta Scale
		1.0 – 1.5	B	
		> 1.5	C	

On the other hand, the validity and reliability of the test were determined using the following method:

a. The construct validity of the test was determined by showing that it is unidimensional. To evaluate unidimensionality, factor analysis was applied.

b. The concurrent validity evidence was secured by examining the relationship between predictors, which is the examinees' test score in the achievement test in Calculus I, and the criterion, which is the grade point average they obtained in their Calculus I course. Pearson Product Moment correlation coefficient is used to examine the relationship between the predictor and the criterion, and in this context the correlation coefficient is referred to as a validity coefficient (Reynolds et al., as cited in Pedrajita, 2009).

c. The content validity of the test was determined by computing a content validity index (CVI), using ratings of scale relevance by content experts. In this study, a 5-point rating agreement scale was used.

d. The internal consistency reliability of the original and the revised test versions was compared using the formula developed by Kuder Richardson, most commonly known as the KR-20. The KR-20 is sensitive to measurement error due to content sampling and is also a measure of item heterogeneity. It is applicable when test items are scored dichotomously, that is, simply right or wrong, as 0 or 1 (Reynolds et al., as cited in Pedrajita, 2009).

Results

I.Detection of Bias Items Using Differential Item Functioning Methods

A. Item Response Theory Models. The results disclosed that the DIF analysis through Transformed Item Difficulty approach detected more DIF items based on the examinees' differences on age. However, the level of sensitivity of the method on the examinees' differences on sex and socio-economic status was very low.The DIF items detected using Transformed Item Difficulty (TID) approach are consolidated in Table 2.

Table 2
Biased Items with Significant DIF across Matched Groups Using TID

Group Comparisons	DIF Items	Total
Age	1,2,3,4,6,7,8,9,11,13,14,16,20,25,27,28,30, 32,33,51,54,56,59,62,65,67,70,72,75,80,8 2,87,91,93,95,96,98	37
Sex	72	1
GPA in English I	6,8,16,27,55	5
Socio-economic Status	None	0
School Type	2,5,12,19,20,23,26,27,31,33,37,38,43,45,5 2,54,55,63,69,70,71,75,76,78,81,86,88,89, 92,95,100	31

Likewise, the Rasch Model DIF analysis also revealed the highest number of DIF items across students' age differences and small number of DIF items across sex and socio-economic status.The results of DIF detection analysis applying Rasch Model (RM) are presented in Table 3.

Table 3
Biased Items with Significant DIF across Matched Groups Using RM

Group Comparisons	DIF Items	Total
Age	1,3,4,7,9,10,11,13,14,15,17,18,19,20,22,23,24, 25,28,29,30,31,32,33,34,35,38,39,40,41,42,43, 44,47,48,49,50,51,52,53,54,55,56,57,58,59,60, 61,62,63,64,65,66,67,68,69,70,71,72,74,75,77, 78,79,80,81,82,85,86,87,88,90,91,92,93,94,95, 96,97,98,99,100	82
Sex	9,24,72,88	4
GPA in English I	7,9,10,11,15,19,22,23,26,33,40,52,53,55,56,61 ,62,67,71,72,74,76,77,78,86,88,90,91,93,97	30
Socio-Economic Status	74,86	2

School Type	1,3,6,7,15,21,23,33,34,35,36,37,41,43,46,47,52,54,58,59,61,62,64,67,69,70,71,72,74,75,88,92,93,95,96,97,99	37

B. Classical Test Theory Models. The findings on the DIF detection analysis using Mantel-Haenszel(MH) Chi-Square Statistic showed that the matched groups based on age differences flaggedmore DIF items. In comparison, results showed that MH Chi-square Statistic was also not very sensitive on detecting biased items in terms of examinees' differences on sex and socio-economic status as compared in the results obtained from using IRT Models. The findings are summarized in Table 4.

Table 4
Biased Items with Significant DIF across the Matched Groups Using MH

Group Comparisons	Identified DIF Items	Total
Age	8,10,11,13,15,16,19,22,34,37,44,52,53,57, 58,61,66,69, 72,74,77,81,96	23
Sex	37,72	2
GPA in English I	6,8,32,55,85,	5
Socio-economic Status	47,86	2
School Type	33,43,44,55,59	5

On the other hand, the Logistic Regression (LR) DIF analysis displayed highest sensitivity in terms of school type differences.In addition, the results showed that matching the examinees across sex and socio-economic status resulted to least number of detected DIF items. Table 5summarizes the DIF items detected using Logistic Regression DIF method.

Table 5
Biased Items with Significant DIF across Matched Groups Using LR

Group Comparisons	DIF Items	Total
Age	1,3,4,5,7,8,12,16,18,21,23,26,29,30,31,32,35,36, 37,38,40,41,45,46,47,48,50,51,60,63,67, 68,73,76,77,78,80,81,82,83,84,89,94,95,96,100	46
Sex	5,27,39,57,59,72,74,88	8
GPA in English I	6,8,16,17,22,26,53,55,61,66,71,73,74,76,79,85,9 0,93,95	19
Socio-economic Status	56,74,86,93,97,99	6
School Type	1,5,9,10,13,17,18,19,20,22,24,26,29,30,32,36,37 ,39,40,43,44,45,49,50,52,53,54,55,60,65,68,69,7 0,71,73,75,76,78,79,81,83,86,87,89,90,92,94,95, 100	49

II. Comparative Analysis on the DIF Detection Models

The comparative analysis of the four DIF methods applied to the 100-item dichotomously scored achievement test in Calculus Ifocused on the number of detected DIF items. The results of the detected DIF items by the four DIF methods are summarized in Table 6. The overall detection was based on the union of the detected items across all the matching variables.

Table 6
Detected DIF Items by the Four DIF Methods across Matched Groups

Matching Variables	Detected DIF Items (%)			
	MH	LR	TID	RM
Age	23	46	37	82
Sex	2	8	1	4
GPA in English I	5	19	5	30
Socio-Economic Status	2	6	0	2
School Type	5	49	31	37
Overall	32	82	60	89

Note. TID – Transformed Item Difficulty; MH – Mantel-Haenszel;LR – Logistic Regression; RM – Rasch Model

III. Validity and Reliability of the Revised Achievement Test

Based on the findings in the DIF analysis using the four DIF methods as well as in the validity and reliability analyses, the achievement test was revised. The revised test was composed of 50 items covering the four subtopics in Calculus I included in the test. It covered concepts on Functions (4 items), Limits and Continuity (13 items), Derivatives (16 items), and Analysis of Functions and their Graphs (17 items). Table 7 presents the final set of items included in the revised test.

Table 7
Items Included in the Revised Achievement Test

Topics	Items	Total
Functions	2,16,18,21	4
Limits and Continuity	26,27,31,33,34,35,36,38,39,41,42, 46,47	13
Derivatives	51,53,54,57,58,60,62,63,66,67,68,70,71, 73,74,75	16
Behaviors of Functions and their Graphs	22,23,24,5,8,10,15,44, 48,49, 77, 80,87, 94, 97, 98, 100	17

Further, the reliability and validity indices of the revised test are presented in Table 8. The table signifies that the revised version of the achievement test in Calculus I is valid, reliable, and a fair test. Thus, the test could be used in evaluating students' performance in Calculus I.

Table 8
Validity and Reliability Test of the Revised Achievement Test

Measures	Coefficient	Description
Construct Validity	0.667	Good
Concurrent Validity	0.159	Significant
Content Validity	0.9793 and 0.8965	Acceptable
Internal Consistency Reliability	0.822	Good

Discussion

I. Detection of Bias Items using Differential Item Functioning Methods

A. Item Response Theory Models. Table 2 shows that the highest number of detected DIF items in the Transformed Item Difficulty Analysis was observed across differences on age. This only indicates that this set of DIF items was not suited to the age level of one group. Hence, these items must be revised or replaced for further improvement of the test.

As gleaned further from the table, matching students in terms of socio-economic status does not detect any potentially biased items. This only show that the performance of the two different groups as based on their socio-economic status does not significantly varies in all the itemsincluded in the test. This finding only indicates that the test items were not bias against these groups. Hence, regardless of their status, the students could have the chance to succeed in all the items. Likewise, differences across gender detected only one DIF item. This also indicates that the students' chance on answering each item correctly in the test were not much affected by their gender differences except on item 72.

On the other hand, Rasch Model DIF analysis detected large number of potentially biased items when the examinees were grouped according to their age. This result states that the students' difference on age wasa great factor that could influence their probability of getting the correct answer to these 82 items. This finding further indicates that the set of items must be revised or replaced in order to suit to the level of ability of the disadvantaged group. Likewise, the finding connotes that the subject must be included in the curriculum of higher year level who are already prepared to take up this course.

Table 3 further indicates that the sensitivity of the Rasch Model in terms of gender and socio-economic status differences was very low. This only shows that thelevel of difficulty of the majority of test items was suited to the level

of ability of the students regardless of their sex and status in life.

B. Classical Test Theory Models. It is visible in Table 4 that the comparisons between groups of students of different age incurred the highest number of potentially biased items in the Mantel-Haenszel Chi-Square Statistic DIF analysis. On the contrary, the analysis detected few DIF items in terms of students' sex and socio-economic status differences. These results coincide with the findings obtained from the IRT Models. Also, this expresses that the sensitivity of the CTT and IRT in terms of age, sex, and socio-economic status differences weresomewhat comparable as based on the result of the test.

On the other hand, Logistic Regression DIF analysis shows that matching the examinees across school type, that is, private versus public HEI's, reveals the highest number of detected DIF items. This finding indicates that this factoralso affects the students' probability of succeeding on the 49 test items flagged with DIF.Hence, majority of the items are bias against school type.Thus, these items must be revised or replaced in order to suit to the capability of the students belonging to the disadvantaged group. Further, this result suggests necessary improvements in the educational system of the affected HEIs.

Moreover, Logistic Regression was also found to be less sensitive in DIF detection in terms of examinees' differences on gender and socio-economic status. The level of sensitivity of this method based on these two matching ability of the students is comparable to the previous three approaches. These results also connote that gender and socio-economic status differences are not contributors to the students' differing performances in the test and did not affect the students' chance of getting the given test items correct. In other words, the test items are not biased against these factors.

II. Comparative Analysis on the DIF Detection Models

The sensitivity of the four DIF methods in detecting DIF items were almost comparable as based on the result of DIF analysis across each matching variables. However, their sensitivity across all the matched groups differed as revealed by Table 6.

Table 6 discloses, that among the four DIF methods, Mantel-Haenszel Chi-Square Statistics detected the least number of DIF items. This implies that this method has the lowest statistical power of detection compared to the three methods. Hence, it is the least sensitive. This result confirms the study of Lopez (2012) which summarizes that the Mantel-Haenszel procedure is a straightforward and adaptable method for detecting DIF but this method has strong limitations which led to the development of other procedures.

On the contrary, Rasch Model appeared to be the most sensitive in the four DIF detection methods for having detected the highest number of items with DIF. This connotes that Rasch Model possesses the highest statistical power of detecting DIF items. Wiberg (2007) states that no matter which method is chosen, it is desirable that the method has high statistical power to detect DIF, that is, having high probability of identifying DIF in an item, while controlling for Type I error, which is the probability of identifying an item as DIF when the item has no DIF.

Moreover, between the two CTT-based methods, Logistic Regression was more sensitive compared to Mantel-Haenszel in detecting potentially biased items. However, it can be observed that the detection power of the Item Response Theory Model, particularly the Rasch Model, is higher than the Classical Test Theory Models. This only strengthens the findings that the latent score is a more precise measure of the ability of the test takers (Wiberg, 2007).

III. Validity and Reliability of the Revised Achievement Test

Construct Validity. The construct validity coefficients revealed that the revised version of the Achievement Test is a good test. Moreover, the results showed that the sampled test items in the revised test represent one dimension.

Concurrent Validity. The concurrent validity coefficient revealed that the revised version of the test obtained a positive relationship between the test score and the grade point average in Calculus I. Moreover, there exists a significant relationship between the two variables. This means that the revised version of the test is valid. However, this difference does not show any significance. This finding supports the results of Roznowski and Reith (1999) and Zumbo (2007) who have reported that DIF has little, if any, impact. Pae and Park (2006) however, reported that DIF may affect the performance on the test.

Content Validity. The content validity indices of the revised version of the test are within the acceptable level. Further, the results indicated that the test was judged valid by the evaluators.

Internal Consistency Reliability. The data in Table 8 revealed that the revised version of the test indexed a reliability coefficient greater than 0.8 which means that the set of test items are good and possess a reliable scale. However, the revised version has lesser reliability coefficient as compared to the original test. It can be observed that the test reliability coefficient obtained decreases when the number of items decreases. This result coincides with the results obtained in the study of Pedrajita (2007) which states that, as more responses on biased items were eliminated, the lower was the internal consistency reliability of the test version. Generally, the two tests were levelled as having a good internal consistency reliability; hence, they are comparable in terms of internal consistency reliability.

Overall, the results of the DIF item elimination on test validity show that the test is valid, reliable, and almost equitable for different types of examinees.

References

Bradley, K., et al. (2009). Constructing and evaluating measures: applications of the Rasch measurement model. *Application of Rasch Measurement,* University of Kentucky, Department of Educational Policy and Evaluation Studies. 131 Taylor Education Building, Lexington, KY 40506-0001.

Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment, 11*(2), 59-76.

Lopez, G. E. (2012).Detection and classification of dif types using parametric and nonparametric methods: A comparison of the IRT-likelihood ratio test, crossing-sibtest, and logistic regression procedures. *Graduate School Theses and Dissertations*. Retrieved from http://scholarcommons.usf.edu/etd/4131.

Madu, B., (2012). Using transformed item difficulty procedure to assess gender-related differential item functioning of mulitple-choice mathematics items administered in nigeria. *Research on Humanities and Social Sciences, 2*(6), 41-56.

Pae T., & Park G. P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing, 23*(4), 475–496.

Pedrajita, J. Q. (2007). Item bias elimination models for test reliability and validity. *Graduate School Theses and Dissertations*. UP Diliman College of Education: Philippines.

Pedrajita, J. Q., & Talisayon, V. M. (2009). Identifying biased test items by differential item functioning analysis using contingency table approaches: a comparative study. *Education Quarterly*, *67*(1), 21-43.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59,* 248–269.

Salubayba, T. (2013). Differential item functioning detection in reading comprehension test using mantel-haenszel, item response theory, and logical data analysis. *International Journal of Social Sciences, 14*(1), 76-82.

Wiberg, M. (2007). Measuring and detecting differential item functioning in criterion-referenced licensing test. *Educational Measurement, 60,* 1-33.

Wood, S. W. (2011). Differential item functioning procedures for polytomous items when examinee sample sizes are small. Retrieved from http://ir.uiowa.edu/ etd/1110

Zumbo, B. D. (2007). Three generations of dif analyses: considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(*2*), 223-233. Lawrence Erlbaum Associates, Inc.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (dif): logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human ResourcesResearch and Evaluation, Department of National Defense.

Zumbo, B. D., & Thomas, D. R. (1997).A measure of effect size for a model-based approach for studying dif. *Working Paper of the Edgeworth Laboratory for Quantitative Behavioral Science*, University of Northern British Columbia: PrinceGeorge, B.C.