Education Journal
2015; 4(4): 139-148
Published online June 18, 2015 (http://www.sciencepublishinggroup.com/j/edu)
doi: 10.11648/j.edu.20150404.11
ISSN: 2327-2600 (Print); ISSN: 2327-2619 (Online)

Using Contingency Table Approaches in Differential Item Functioning Analysis: A Comparison

Jose Quito Pedrajita

Educational Research and Evaluation Department, Division of Educational Leadership and Professional Services, University of the Philippines College of Education, Diliman, Quezon City, Philippines

Email address:
josepedrajita@gmail.com

To cite this article:
Jose Quito Pedrajita. Using Contingency Table Approaches in Differential Item Functioning Analysis: A Comparison. *Education Journal*. Vol. 4, No. 4, 2015, pp. 139-148. doi: 10.11648/j.edu.20150404.11

Abstract: This study provides a demonstration of differential item functioning (DIF) analysis. It made use of test scores of 200 junior high school students on a Chemistry Achievement Test, a measure tested for its psychometric properties. One hundred students came from a public school, while the other 100 were private school examinees; one hundred students were males and the other 100 were females; and 95 students were of low ability and 105 students were of high ability based on their English II grades. Four contingency table approaches, the Chi-Square, Distractor Response Analysis, Logistic Regression and the Mantel-Haenszel Statistic, were applied in the DIF analysis to identify test items indicating bias between examinees matched on school type, gender, and English ability. Thereafter, the results for the four approaches were compared. The findings revealed the presence of items indicating school type-, gender-, and English ability-based DIF. There was a high degree of correspondence between the Logistic Regression and the Mantel-Haenszel Statistic in identifying potentially biased test items.

Keywords: Contingency Table Approaches, Differential Item Functioning, Differential Item Functioning Analysis, Item Bias

1. Introduction

A critical step in the development and adaptation of psychological assessment instruments is ensuring that no individual or group responding to the instrument is disadvantaged in any way (Kanjee, 2007). For instance, in an achievement test, equally able examinees in terms of ability and drawn from the same population but belonging to different subgroups such as male or female, should have the same chance of correctly responding to an item. Biased test items are those that differentially inhibit individuals from showing their true abilities. They are said to be displaying differential item functioning (DIF) which according to Reynolds (2006) systematically underestimates or overestimates the value of the variable the items are designed to measure.

Test fairness is a crucial issue in testing. A test that is not fair is a biased test. The process for developing instruments that are fair for all test takers requires the removal or revision of potentially biased items. In practice, this implies that before any instrument is ready for use, all biased items are first detected, and either eliminated or revised. Questions of test bias are closely related to questions of test validity. A test possesses validity if it measures what it purports to and invalidity if it does not. Bias is a kind of invalidity that arises relative to groups.

Validity is an essential requirement of all tests. A valid test produces outcomes that are based only on the trait being measured rather than irrelevant characteristics (Fidalgo, 2011). When test scores depend on irrelevant characteristics such as group membership (i.e., gender, age social; status) then the test is considered as potentially biased. Bias is a technical term that simply refers to "the consistent distortion of a statistic" (Osterlind, 1983) and does not necessarily imply test unfairness.

One way to investigate bias at the item level is through differential item functioning (DIF) analysis. DIF analysis is a means of statistically identifying unexpected differences in performance across matched groups of examinees. It compares the performance of matched majority (or reference) and minority (or focal) group examinees.

Differential item functioning (DIF) is said to be present in a test item when, despite controls for overall test performance, examinees from different groups have a different probability

of answering an item correctly or when examinees from two subpopulations with the same trait level have different expected scores on the same item (Camilli and Shepard, 1994; Kamata and Vaughn, 2004; Penfield & Camilli, 2007; Roussoss & Stout, 2004; Kristjansson, Aylesworth, McDowell, & Zumbo, 2005). Thus, an item that exhibits DIF may or may not be biased for or against any group (Kanjee, 2007). DIF may be attributed to item bias but may also reflect performance differences that the test is designed to measure (Camilli and Shepard, 1994).

Bias is not the mere presence of a score difference between two groups. In test items, bias is the presence of a systematic error in measurement (Camilli and Shepard, 1994). Items may be judged relatively more or less difficult for a particular group by comparison with the performance of another group drawn from the same population.

In test theory, the chance of an examinee correctly responding to an item is called the probability of success. A test item is said to be unbiased when the probability for success on the item is the same for equally able examinees of the same population regardless of their group membership (Osterlind, 1983).

In this study, four contingency table approaches were applied to a researcher-constructed and validated achievement test in Chemistry to detect differential item functioning. It looked into items that function differently between public and private, male and female, and low and high English ability examinees through DIF analysis. It also looked into the agreement among the DIF approaches in identifying items which function differently between the matched examinees.

It is important to note, however, that 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 (Camili & Sheppard, 1994). Logical analysis is meant to discover detectable patterns of DIF or common characteristics of individual items. However, this study focused only in the statistical analysis of bias or DIF.

2. Method

This study employed the descriptive-comparative research design. Three reference/focal group combination were used in the *differential item functioning* analysis. The first reference/focal group combination was between the 100 public and the 100 private school examinees. The second was between the 100 male and the 100 female examinees. And the third was between the 95 low and the 105 high ability examinees. The examinees were third year high school students taken from the top, middle, and lower class sections of a selected public and private school. For each matched groups the total number of examinees adds up to 200, which was the total sample in this study. All examinees were matched by sections and total score.

The preparation of the Chemistry Achievement Test involved the following steps: (1) development of a Table of Specifications; (2) consultation with adviser/experts; (3) generation of an item pool; (4) review of the initial item pool by experts; (5) field-testing; and (6) item analysis and test revision.

The data gathering procedures involved: (1) administration of the test to the intact classes; and (2) checking and scoring the test. The analysis of data involved (a) assignment of examinees' test papers to the three comparison groups matched by section and total score; (b) organizing data for every item into a three-way contingency table; (c) encoding data in the Statistical Analysis System (SAS) computer program; (d) analysis for detecting and testing for differential item functioning for each comparison group.

The four *contingency table* (CT) approaches applied in the differential item functioning analysis were the Chi-Square, Distractor Response Analysis, Logistic Regression, and the Mantel-Haenszel Statistic. These methods were chosen because they can be applied to small sample sizes. In fact, smaller samples are required for the CT methods for a number of reasons. *First*, total ability for a particular examinee is estimated by that person's score on the entire test. Total test scores yield a valid indicator of ability. *Second*, no provision is made for guessing; the assumption is that the guessing parameter is equal for two groups on each item. *Finally*, no provision is made for variation in the discriminating power of test items; the assumption is that for each item the discrimination parameter is the same for both the focal and reference groups.

The *Chi-Square* (X^2) approach to the identification of item bias examines the likelihood or probability of test takers from different groups with the same ability levels correctly responding to an item. The hypothesis under test is that *there is no significant difference in proportions attaining a correct response across total score categories on the test items between the reference and the focal group.*

The *Distractor Response Analysis* (DRA) examines the incorrect alternatives to a test item for differences in patterns of response among different subgroups of a population. It determines the significance of the differences among two or more group's response frequencies in the discrete categories of question distracters. The hypothesis under test is that *there is no significant difference in proportions selecting distracters on the test items between the reference and the focal group.*

The *Logistic Regression* (LR) is a kind of regression analysis often used when the dependent variable is dichotomous and scored 0 or 1. It is usually used for predicting whether something will happen or not. Independent variables may be categorical or continuous. The hypothesis under test is that *for two groups at level j, the population value is zero for either the difference between the proportions correct or the log odds ratio on the test items between the reference and the focal group*.

In the LR analysis between the matched examinees, the independent variables were: *score interval* and *school type* for the public/private matched examinees; *score interval* and *sex* for the male/female matched examinees; and *score*

interval and *ability* for the low/high ability matched examinees. The dependent variable or logit for each of the matched examinees is *the odds or likelihood of getting the item right.* A significant *score interval,* in each of the matched examinees, indicates that examinees with higher total score tend to score better in the examination. A significant *school type, gender, and ability,* indicates that *the odds of getting an item right* are different between the public/private, male/female, and low/high ability matched examinees, respectively.

The *Mantel-Haenszel Statistic* (MH) is a non-parametric contingency table procedure commonly used to perform statistical test for uniform DIF. *Uniform DIF* refers to one group having a constant advantage at each level of ability or when the magnitude of DIF is the same across all trait levels. Whereas, *non-uniform DIF* refers to the relative advantage of one group at one ability level, but a disadvantage at another or when the magnitude of DIF is not consistent across trait levels (Camilli and Shepard, 1994; Kamata and Vaughn, 2004). MH yields a chi square test with one degree of freedom to test the null hypothesis that *there is no significant relationship between group membership and test performance on the test items between the reference and the focal group.* MH uses an internal matching variable, total test score, when evaluating the suspect item, to ensure that the examinees at each score level are comparable.

In this paper, all tests of hypotheses were carried out at the 0.05 level of significance. *Statistical bias* or *DIF* is inferred if the probability associated with the obtained chi-square value is less than the set alpha level of 0.05 with one degree of freedom. *School type-*, *gender-*, and *ability-based DIF* refer to the differing probabilities of success on an item between the public and the private, the male and the female, and the low and the high English ability examinees, respectively.

The agreement of any two, three or all of the DIF methods was indicated by their obtained measure of bias (chi-square value). If any two, three or all of the four approaches similarly obtained a statistically significant chi square value on an item or number of items, such methods were in agreement. If not, there is disagreement.

3. Results and Discussion

3.1. Differential Item Functioning Analysis

3.1.1. School Type-Based DIF

Table 1 shows the differentially functioning items between the public and private school matched examinees.

Table 1. *Differentially Functioning Items Detected in the Public/Private Matched Examinees.*

Items	Concept/Skills Measured	Indicates DIF against	Identified by
1	gas property illustrated by garbage smell entering the house	Private	X^2 DRA LR MH
2	element with Latin name "aurum"	Public	LR MH
3	chemical bond which held together two atoms in a molecule by the transfer of an electron from one atom to the other	Private	X^2 DRA LR MH
5	Filipino scientist who pioneered in the use of biogas/ biomass as a source of energy	Private	X^2 DRA LR
8	definition of valence electrons	Public	DRA LR MH
9	description of dialysis	Private	X^2 DRA LR MH
10	volume of a cube	Private	LR MH
13	new pressure of the gas when the volume is compressed to a smaller quantity	Public	X^2 DRA LR MH
14	problem on Boyle's Law	Public	LR MH
16	how the chemical and molecular formula of sodium sulfate is correctly written	Public	X^2 DRA LR MH
19	solving for the molar mass of $Fe_2 O_3$	Private	X^2 DRA LR MH
21	the mass of oxygen in sulfur trioxide if the ratio of sulfur to oxygen is 2:3 with sulfur having a mass of 6 grams	Private	DRA LR MH
22	volume conversion	Public	X^2 DRA LR MH
26	indicators of chemical change	Private	LR MH
30	correct position of Chlorine in the periodic table	Private	X^2 DRA LR MH
31	indicator of a balanced chemical equation	Private	X^2
32	which chemical equation is balanced	Public	LR MH
33	identify the reactants in the given chemical equation	Private	X^2 DRA LR MH
35	identify which principle is true of different substances having an equal number of moles	Private	DRA
36	classification of a solution which changes red litmus paper to blue	Public	DRA LR MH
37	factors which increases the solubility of a solute	Public	X^2 LR MH
40	evidences of chemical change	Public	LR MH
41	laws which govern changes in matter	Public	DRA LR MH
43	properties of gases	Private	DRA
46	components of a solution	Public	MH
47	strategy which is most probable in proving the given hypothesis in the given experiment	Private	X^2 DRA LR MH
50	factor which causes the nails to rust	Private	DRA

The Chi-Square analysis identified 13 items indicating DIF between the public and private school examinees. Nine of which, items 1, 3, 5, 9, 19, 30, 31, 33, and 47 were pointing bias against the private school examinees. Whereas, four

items, items 13, 16, 22, and 37 were showing bias against the public school examinees.

The DRA revealed 18 items which indicate DIF between the public and private school examinees. These were items 1, 3, 5, 8, 9, 13, 16, 19, 21, 22, 30, 33, 35, 36, 41, 43, 47, and 50. Twelve of which, items 1, 3, 5, 9, 19, 21, 30, 33, 35, 43, 47, and 50 were showing bias against the private school examinees. Whereas, six, items 8, 13, 16, 22, 36, and 41 were indicating bias against the public school examinees.

The LR analysis identified 22 items which indicate DIF between the public and the private school examinees. These were items 1, 2, 3, 5, 8, 9, 10, 13, 14, 16, 19, 21, 22, 26, 30, 32, 33, 36, 37, 40, 41, and 47. Of which, eleven items, 1, 3, 5, 9, 10, 19, 21, 26, 30, 33, and 47 were potentially biased against the private school examinees. In each of these items, the odds of getting an item right favored the public school examinees. Whereas, the other eleven items, 2, 8, 13, 14, 16, 22, 32, 36, 37, 40, and 41 were indicating bias against the public school examinees. In each of these items, the odds of getting an item right favored the private school examinees.

The MH analysis between the public and the private school examinees showed that 22 of the 50 items displayed DIF. Of these 22 items, ten favored the public school examinees. They were items 1, 3, 9, 10, 19, 21, 26, 30, 33, and 47. Each of these ten items obtained a significant MH chi square value and positive log odds ratio signifying DIF in favor of the public school examinees. Whereas, twelve items favored the private school examinees, namely, items 2, 8, 13, 14, 16, 22, 32, 36, 37, 40, 41, and 46. Each of these twelve items obtained a significant MH chi square value and a negative log odds ratio indicative of DIF in favor of the private school examinees.

3.1.2. Gender-based DIF

Table 2 shows the differentially functioning items detected in the male and female matched examinees

The chi-square analysis reveals that only one item, item 17, was found indicating bias between the male and the female examinees. That is, this item favored the male examinees.

Table 2. *Differentially Functioning Items Detected in the Male/Female Matched Examinees.*

Items	Concept/Skills Measured	Indicates DIF against	Identified by
1	gas property illustrated by garbage smell entering the house	Male	LR MH
3	chemical bond which held together two atoms in a molecule by the transfer of an electron from one atom to the other	Male	LR MH
17	electron configuration of the element Sodium	Female	X^2 LR MH
27	options which illustrates the compressibility of gases	Female	DRA LR MH
34	definition of reaction reversibility	Female	LR MH
42	principles of Kinetic Molecular Theory	Male	DRA LR MH
47	strategy which is most probable in proving the given hypothesis in the given experiment	Male	LR MH

The DRA showed 2 items indicating bias between the male and the female examinees. They were items 27 and 42. Item 27 was showing bias against the female examinees. The female examinees obtained a large number of responses on the incorrect options, indicating less familiarity with the concept reflected in the item. Hence, this item favored the male examinees. Conversely, item 42 was indicating bias against the male examinees. The male examinees obtained a differentially large number of responses in the incorrect options. Hence, this item favored the female examinees.

The LR analysis identified 7 items which indicate bias between the male and the female examinees. These were items 1, 3, 17, 27, 34, 42, and 47. Three of which, items 17, 27, and 34, were showing bias against the female examinees. Whereas, items 1, 3, 42, and 47 were indicating bias against the male examinees. These items favored the female examinees.

The MH analysis between the male and the female examinees revealed that 7 of the 50 items displayed DIF. Of the seven DIF items, three favored the male examinees. These were items 17, 27, and 34. Each of these three items obtained a significant MH chi square value and a positive log odds ratio signifying DIF in favor of the male examinees. Whereas, items 1, 3, 42, and 47 favored the female examinees. These four items obtained a significant MH chi square value and a negative log odds ratio, indicative of DIF in favor of the female examinees.

The succeeding citations from literature showed test performance differences between male and female, girls and boys in the different areas and skills the various tests measured. Gender is the most studied variable in DIF analysis.

Gierl's (1999) study evaluated the effects of differential item functioning between males and females on the Alberta Education Social Studies 30 Diploma Examination. The multiple choice section of the examination contained 70 items, each with four options. The results from the statistical analysis indicate that the majority of multiple choice items do not display differential item functioning. Using the three-tiered ratings, 65 of the 70 items displayed negligible effects, five of the 70 items displayed moderate effects, and none of the items displayed large effects. Of the five items with moderate DIF, three favored males and two favored females. This indicates that the test contained items that functioned differently for males and females.

Gender differences usually exist in test item analysis. On tests of spatial skill requiring visualization and imagery, men and boys tend to score higher than do women and girls. On

tests that involve written language and tests of simple psychomotor speed (such as the rapid copying of symbols or digits), women and girls tend to score higher than men and boys (Reynolds, 2006).

Performance differences between males and females on a test may be a product of environment. In the study of SAT, the balancing of the verbal content was made so that both sexes received similar average verbal scores on the test. Girls do significantly better on questions that are neutral or have female actors. The item showing the most favoritism to men was a math question that asked about a "basketball team won/loss record", more percentage of boys than girls answered that question correctly (Wolf & Phyllis, 1990).

DIF in favor of each gender corresponded to traditional sex-role stereotypes; males perform better on "masculine" items, whereas females are advantaged on "feminine" questions (Le, 1999).

Barnett and Ercikan (2006) confirmed that problem solving as a content area was a source of gender DIF in favor of boys when an item is presented in the form of a story problem or when the problems are non-context specific. Items in Geometry were not found to be a source of gender DIF. All of the high cognitive level items favored boys. High levels of DIF were detected in favor of girls on the bundle of computation items in which no equations were provided in the question.

3.1.3. English Ability-based DIF

Table 3 shows the differentially functioning items detected in the low and high ability matched examinees.

The Chi-Square analysis between the low and high ability examinees identified seven DIF items, namely items 2, 3, 6, 8, 13, 19, and 30, which were all showing bias against the low ability examinees. That is, these items favored the high ability examinees. None of the items, however, were indicating bias against the high ability examinees.

The DRA revealed eleven items which indicate bias between the low and the high ability examinees. These were items 3, 6, 7, 8, 13, 15, 19, 22, 30, 48, and 50. All of them were indicating bias against the low ability examinees. In each of these items, one, two or all of the three incorrect options had obtained large number of responses from the low ability examinees. Such incorrect options seem plausible for the said examinees. Thus, these items favored the high ability examinees.

The LR analysis identified thirteen items which indicate bias between the low and the high ability examinees. These were items 2, 3, 6, 8, 13, 19, 22, 29, 36, 38, 45, 48, and 50. Three of which, items 2, 9, 38, and 45, were showing bias against the high ability examinees. These items favored the low ability examinees which scored higher in the items. Whereas, ten items, 2, 3, 6, 8, 13, 19, 22, 36, 48, and 50 were indicating bias against the low ability examinees. These items favored the high ability examinees which scored higher in the items.

Table 3. *Differentially Functioning Items Detected in the Low/High Ability Matched Examinees.*

Items	Concept/Skills Measured	Indicates DIF against	Identified by
2	element with Latin name "aurum"	Low	X^2 LR MH
3	chemical bond which held together two atoms in a molecule by the transfer of an electron from one atom to the other	Low	X^2 DRA LR MH
6	scope of chemistry	Low	X^2 DRA LR MH
7	property of gases that best describes the foul odor of a nearby garbage dump	Low	DRA
8	correct definition of valence electrons	Low	X^2 DRA LR MH
13	new pressure of the gas when the volume is compressed to a smaller quantity	Low	X^2 DRA LR MH
15	problem-solving on Charles' Law	Low	DRA
17	electron configuration of the element Sodium	Low	MH
19	solving for the molar mass of Fe_2 O_2	Low	X^2 DRA LR MH
22	volume conversion	Low	DRA LR MH
29	valence electrons of the Chlorine atoms	High	LR MH
30	correct position of Chlorine in the periodic table	Low	X^2 DRA
36	classification of a solution which changes red litmus paper to blue	Low	LR MH
38	in which solution water is a solute	High	LR MH
45	in which situation the process of oxidation is common	High	LR MH
48	correct formula in solving for the new volume of the gas	Low	DRA LR MH
50	factor which causes the nails to rust	Low	DRA LR MH

In the MH analysis between the low and the high ability examinees, 14 of the 50 items indicate bias. Of the fourteen DIF items, three favored the low ability examinees. They were items 29, 38, and 45. These three items obtained a significant MH chi square value and a positive log odds ratio signifying DIF in favor of the low ability examinees. Whereas, eleven items favors the high ability examinees. They were items 2, 3, 6, 8, 13, 17, 19, 22, 36, 48, and 50. These items obtained a significant MH chi square value and a negative log odds ratio indicative of DIF in favor of the high ability group.

Thus, the null hypothesis under test for each of the DIF approaches for the three reference-focal group combinations is therefore rejected in favor of their alternative hypothesis.

3.2. Agreement of the DIF Methods in Detecting Differential Item Functioning

Table 4 shows the agreement between and among the approaches on DIF items detected. The upper column contains the potentially biased items against the *private*, *female* and *high ability* examinees, while, the lower column contains the potentially biased items against the *public*, *male*, and *low ability* examinees.

Table 4. *Agreement of DIF Methods in Detecting Differential Item Functioning.*

	Chi Square			**Distractor Response Analysis**			**Logistic Regression**			**Mantel-Haenszel Statistic**		
	School Type	**Gender**	**Ability**	**School Type**	**Gender**	**Ability**	**School Type**	**Gender**	**Ability**	**School Type**	**Gender**	**Ability**
	1			1			1			1		
	3			3			3			3		
	5			5			5					
	9			9			9			9		
							10			10		
		17						17			17	
	19			19			19			19		
				21			21			21		
							26			26		
					27			27			27	
									29			29
	30			30			30			30		
	31											
	33			33			33			33		
								34			34	
				35								
									38			38
				43								
									45			45
	47			47			47			47		
				50								
								1			1	
			2				2		2	2		2
			3			3		3	3		3	3
			6			6			6			6
						7						
			8	8		8	8		8	8		8
	13		13	13		13	13		13	13		13
							14			14		
						15						
	16			16			16			16		
												17
			19			19			19			19
	22			22		22	22		22	22		22
			30			30						
							32			32		
			36				36		36	36		36
	37						37			37		
							40			40		
				41			41			41		
					42			42			42	
										46		
								47			47	
						48			48			48
						50			50			50
Total	13	1	7	18	2	11	22	7	13	22	7	14

The DIF analysis between the public and private school examinees reveals that there were items that were singly or identically identified by one, two, three, or all of the approaches.

Ten items were identically identified by the four methods. Seven of which, items 1, 3, 9, 19, 30, 33, and 47 were indicating bias against the private school examinees. These items have indices of difficulty within 0.5 to 0.78. That is, these difficulty indices indicate that these were relatively easy items, being above the 0.5 level of difficulty. However, item 1 was also identically identified in the LR and MH analyses as indicating bias against the male examinees. Item 3 was also identically identified by the four methods as potentially biased against the low ability examinees and further identified in both the LR and MH analyses as indicating bias against the male examinees. Moreover, item

19 was also identically identified by the four methods as potentially biased against the low ability examinees. Item 30 was also identified as indicating bias against the low ability examinees in both the X^2 and DRA. Still, item 47 was also identified by both the LR and MH Statistic as showing bias against the male examinees. Whereas, three, items 13, 16, and 22 were signaling bias against the public school examinees. They have indices of difficulty which ranged from 0.16 to 0.36. These difficulty indices indicate that these items are relatively difficult items, being lower than the 0.5 level of difficulty. Thus, the relatively easy items were indicating bias against the private school examinees and the relatively difficult items were showing bias against the public school examinees. However, item 13 was also identically identified by the four methods as potentially biased against the low ability examinees. Moreover, item 22 was likewise identically identified in the DRA, LR, and MH analyses as signaling bias against the low ability examinees.

Four items were identically identified by the DRA, LR, and MH statistic. One of which, item 21, was pointing bias against the private school examinees. Its difficulty index of 0.78 indicates that it was an easy item. Conversely, items 8, 36, and 41 were showing bias against the public school examinees. Their difficulty indices ranged from 0.28 to 0.54 which means that these items were relatively difficult though they were within the middle range or optimum difficulty level. Moreover, item 8 was identically identified by the four methods as indicating bias against the low ability examinees. In addition, item 36 was also identified by both LR and MH Statistic as showing bias against the low ability examinees.

Only one item, item 5, was identically identified as signaling bias against the private school examinees in the X^2, DRA, and LR analyses. It has difficulty index of 0.76, indicating that it is an easy item.

Another lone item, item 37, was identically identified in the X^2, LR, and MH analyses as showing bias against the public school examinees. It has difficulty index of 0.38, indicating that it is relatively a difficult item.

Six items were identically identified in the LR and MH analyses. Two of which, items 10 and 26 were indicating bias against the private school examinees. They have difficulty index of 0.7 and 0.84, respectively, indicating that these were relatively easy items. Whereas, four items, items 2, 14, 32, and 40 were potentially biased against the public school examinees. Their difficulty indices were 0.21, 0.34, 0.46, and 0.79, respectively. These difficulty indices indicate that these items were relatively difficult, with the exception of item 2. However, item 2 was also identically identified as showing bias against the low ability examinees in the X^2, LR, and MH analyses.

Three items, items 35, 43, and 50 were each identified only in the DRA as showing bias against the private school examinees. Their difficulty indices were 0.33, 0.64, and 0.66, respectively. Though all of them belong to the middle range of difficulty, items 43 and 50 were relatively easier than item 35. However, item 50 was also commonly identified in the DRA, LR, and MH analyses as indicating bias against the low ability examinees.

A lone item, item 31, was singly identified only in the X^2 analysis as showing bias against the private school examinees. It has difficulty index of .64 indicating that it was a relatively easy item, being above the 0.5 level of difficulty.

Another single item, item 46, was identified only by the MH Statistic as signaling bias against the public school examinees. It has difficulty index of 0.73 indicating that it is relatively easy item, being above the 0.5 level of difficulty. Though the MH method did not obtain a statistically significant chi square value, the item nevertheless falls on the C category of items which was considered biased because of large DIF effect, indicated by a delta-MH greater than 1.5 in magnitude.

A clear pattern in the analysis shows that potentially biased items against the private school examinees were relatively easier items, mostly within the middle and upper ranges of difficulty levels. Whereas, potentially biased items against the public school examinees were relatively difficult items, mostly falling within the middle and lower ranges of difficulty levels.

The LR and the MH Statistic approaches yielded very similar results. Both identified 22 potentially biased items, 21 of which were identical items, except for item 5 for the LR and item 46 for the MH Statistic.

The DIF analysis between the male and the female examinees indicates that there were also items which were identified singly or identically by one, two, three, or all of the four DIF approaches.

Item 17 was commonly identified in the X^2, LR, and MH analyses as showing bias against the female examinees. It has difficulty index of 0.86, indicating that it is a very easy item. Moreover, item 17 was also identified as indicating bias against the low ability examinees solely by the MH Statistic. Although the MH analysis did not obtain a significant chi square value, its delta-MH, being higher than 1.5, reveals that it was a statistically biased item.

Items 27 and 42 were identically identified in the DRA, LR, and MH analyses. Item 27 was showing bias against the female examinees. It has difficulty index of 0.58, indicating that it is a relatively easier item. On the other hand, item 42 was indicating bias against the male examinees. It has difficulty index of 0.42. Compared to item 27, this is a relatively difficult item.

Both the LR and the MH statistic commonly identified items 34 and 47. Item 34 was pointing bias against the female examinees. Its difficulty index is 0.18, indicating that it is a difficult item. Conversely, item 47 were showing bias against the male examinees. Its difficulty index was 0.78. Compared to item 34, this is a relatively easy item.

The analysis revealed that the LR and the MH Statistic were most similar among the four DIF approaches. Each identified identical and similar number of items.

Likewise, the DIF analysis between the low and the high ability examinees showed that there were items which were identified solely or identically by one, two, three, or all of the four DIF methods.

Item 6 was commonly identified by the four approaches as showing bias against the low ability examinees. Items 6 have difficulty index of 0.79, indicating that it is a relatively easy items.

Item 48 was also identically identified in the DRA, LR, and MH analyses as indicating bias against the low ability examinees. Its difficulty index of 0.7 indicates that it is an easier item.

Items 29, 38, and 45 were commonly identified as potentially biased against the high ability examinees by the LR and the MH Statistic. Their difficulty indices were 0.38, 0.41 and 0.5, respectively. That is, these items were of optimum difficulty, being at the middle range of difficulty indices.

Items 7 and 15 were identified solely in the DRA. Their difficulty indices were 0.74 and 0.56, respectively, indicating relatively easier items because, though in the middle range of difficulty ranges, their difficulty indices were above the 0.5 difficulty index.

A closer scrutiny of the potentially biased items against the low ability examinees shows that the difficulty indices of all these items belong to the middle up to the upper ranges of difficulty levels. That is, these items have difficulty indices ranging from optimum difficulty to very easy, mostly higher than the 0.5 index of difficulty, except items 13, 22, and 36. Whereas, the potentially biased items against the high ability examinees have difficulty indices within the middle range or optimum difficulty level and less than the 0.5 level of difficulty. Hence, the pattern of bias generally points toward the low ability examinees.

The LR and the MH analyses for the low/high ability examinees yielded very similar results. Each identified 13 identical biased items, except for the MH chi square which identified item 17, giving it an extra one item more than the LR.

Overall, there were items that were identified singly by one as well as identically by two, three, or all of the DIF methods in the three reference-focal group combinations.

Previous studies have also examined the agreement among several dichotomous DIF techniques in detecting potentially biased items. For instance, Wiberg (2009) compared DIF detection and effect sizes of log linear modeling, logistic regression, and Mantel-Haenszel procedures using driving test. On the other hand, Wang & Lane (1996) implemented a DIF analysis on a cognitive assessment tool for mathematics using logistic regression, discriminant function analysis, and HW3 method.

Several studies (MJ Navas-Ara & Gomez-Benito, J., 2002; Nijenhuis et al., 2004; Sheppard, Han, Colarell, Dai, & King, 2006; Stoneberg, 2004; Wolf & Phyllis, 1990) advocate the use of Mantel-Haenszel Chi-Square in detecting DIF. The procedure has been found simple as it does not require highly specialized software.

Several bias-detection techniques were analyzed in the study about the effects of ability scale purification on the Identification of Differential Item Functioning using the Mantel-Haenszel statistic as among the DIF techniques (MJ Navas-Ara, 2002). Results showed that purifying the ability scale improved item bias detection greatly, providing rates of correct identification close to 100% with all these techniques. IRT-based indices showed the greatest improvement. But overall results suggest that Mantel Haenszel has the greatest advantages as a DIF detection technique, since it is the simplest and it provides the best results without purification.

Investigation of DIF was examined across sex and two racial groups in the Hogan Personality Inventory (Sheppard, et. al, 2006). The study used the Mantel-Haenszel chi-square method to detect DIF. Items displaying DIF were slightly more cohesive for sex than for race.

The influence of cultural background on the intellectual performance of children from immigrant groups using the RAKIT intelligence test for immigrant children was examined in the study of Nijenhuis (2004). The Mantel-Haenszel DIF method detected biased items against Mexican-Americans as they score lower on the intelligence test than do whites.

The Mantel-Haenszel Chi Square test and Simultaneous Item Bias Test (SIBTEST) were used to detect Gender-Based Differential Item Functioning (DIF) in the Spring 2003 Idaho Standards Achievement Tests (Stoneberg, 2004). The proportion of items exhibiting gender-based DIF ranged from seven percent for the Grade 4 reading test to 37 percent for the Grade 10 mathematics test.

3.3. Choice of DIF Method

In deciding which DIF approach to use, it is appropriate to choose method which is most valid. Valid methods may be very sensitive and may have a very high detection rate in identifying biased test items. But then, it is better for test development for it could identify all items which are possibly biased, and then to eliminate, replaced or revised such biased items in order to purify and maintain the measurement qualities of the test.

On the other hand, if methods which may not be so sensitive and with a very low detection rate are used, some items which could be possibly biased may be not identified and may remain part of the test content, thereby, still affecting and contaminating the validity and reliability of the test.

DIF approaches with high detection rate are preferable over those with low detection rate in purifying assessment instrument, indicating that test items should be free of bias. Another factor to consider in the choice of DIF approaches are external evidences of validity. Some external evidence of validity for a DIF technique would be a demonstration that the technique is not selecting item at random and the results obtained with different approaches tend to agree. In this study, the LR and MH procedures had demonstrated such external validity evidence. The two procedures result in similar number of items (and similar items) being identified.

4. Conclusion

The results of the differential item functioning analysis

showed that there were statistically biased test items between the public and the private, the male and the female, and the low and the high ability examinees. A clear pattern shows that the potentially biased items against the private school examinees were relatively easier items, mostly having difficulty indices of 0.5 and above. Whereas, potentially biased items against the public school examinees were relatively difficult, mostly within the difficulty ranges of 0.5 and below. Overall, it appears that students from public schools performed better than those from private schools in the Chemistry Achievement Test. The test items were generally fair between the male and the female examinees. Potentially biased items against the low ability examinees shows that the difficulty indices of all these items belong to the middle up to the upper ranges of difficulty levels. That is, these items have difficulty indices ranging from optimum difficulty to very easy, mostly higher than the 0.5 index of difficulty. Whereas, the potentially biased items against the high ability examinees have difficulty indices within the middle range or optimum difficulty level and less than the 0.5 level of difficulty. Hence, the pattern of potential bias points mostly against the low ability examinees.

There was agreement and disagreement among the DIF approaches in the identity and number of items identified. There were items which were identically identified (a) by the four methods, (b) by three of the four methods, (c) by two of the four methods, and (d) by a single method. If any two, three or all of the four DIF approaches similarly obtained a statistically significant chi square value on an item or groups of items, such methods were in agreement. If not, there is disagreement.

The Logistic Regression and the Mantel-Haenszel Statistic yielded very similar results with respect to uniform differential item functioning (DIF). The two procedures result in similar number and identity of items being identified. Hence, there is high degree of correspondence between these two procedures.

Recommendations

In this study, only statistical analysis of DIF was conducted. It is recommended that the DIF items should be subjected to content review by curriculum specialist in order to address the possible sources or causes of DIF. Quantitative outcomes of DIF analysis should always be supported by qualitative reviews to ensure that only items with "explainable sources of bias" are considered for removal or revision. Focus group discussion with the experts may be conducted after they have individually explained the sources of DIF/bias to discuss their viewpoints and to come up with a consensus as to which items are actually biased and the reasons for such bias.

Test experts and developers should consider using contingency table (CT) approaches, preferably the LR and MH approaches in DIF detection. The two procedures are viable in DIF detection and result in similar number of items (and similar items) being identified and has a very high detection rate.

DIF detection could be useful for DIF/bias correction, which means that identified DIF/biased items should be revised or replaced. DIF/bias correction could make differentially functioning items between groups of interest be more valid, reliable, and fair. Bias correction could maintain or improve the measurement qualities of a test such as its content validity, concurrent validity, and internal consistency reliability.

DIF/biased items must either be revised or replaced since its elimination and non-replacement lessens the number of items in a test. The lesser the number of items, the smaller the content validity, concurrent validity, and internal consistency becomes.

In this study, matching was done by conditioning simultaneously on test score, and a categorical variable, namely, total score and school type for the public/private comparison group, total score and sex for the male/female comparison group, and total score and English ability for the low/high ability comparison group. In connection with the above-mentioned conditioning, it is also recommended that a study be conducted by incorporating more than two or multiple ability estimate into a DIF/item bias analysis. That is, matching should be 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. Other grouping variables may be considered in future DIF analyses that are deemed to influence performance in the test holding ability constant.

Future studies should consider focusing on comparative study of Item Response Theory (IRT) models and Contingency Table (CT) approaches in DIF detection. The study suggests that researchers conduct similar DIF studies using both IRT and LR or MH in other tests beside achievement tests. Several authors (Baker & Kim, 2004; Embretson & Reise, 2000; Hambleton, 1991) advocate the use of IRT in DIF detection which for them is more promising than the CTT DIF techniques.

Educational institutions, educational evaluators, and test experts and developers should consider giving increasing attention to equity of test scores for various groups or subgroups of examinees. Test equity can be achieved by ensuring that a test measures only construct-relevant differences between groups or subgroups of examinees. To achieve test equity among groups or subgroups of examinees, DIF testing must be conducted especially for very important tests like entrance examination and professional licensure examination.

One of the objectives of this paper is on detecting DIF items. However, it is also recommended that further studies be conducted to go beyond detecting DIF items and obtain additional information about DIF. Some items may show larger magnitude of DIF, while some others show relatively small magnitude of DIF. In such a situation, it is of interest to investigate sources of such variation. The impact of high occurrence of DIF needs further investigation.

References

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

[2] Barnett, S. & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. Applied Measurement in Education. 19(4), 289-304.

[3] Camilli, G. and Shepard, L. (1994). Methods for Identifying Biased Test Items. Volume 4, Sage Publications, Inc., California.

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

[5] Fidalgo, A. M. (2011). A new approach for differential item functioning detection using Mantel-Haenszel Methods. The GMHDIF Program. The Spanish Journal of Psychology, 14:2.

[6] Gierl, M. J. (1999). Differential item functioning on the Alberta Education Social Studies 30 Diploma Examination. Canadian Social Studies, Vol. 33, No. 2.

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

[8] Kamata, A. and Vaughn B. (2004). An introduction to differential item functioning analysis. Learning Disabilities: A Contemporary Journal 2(2), 49 – 69.

[9] Kanjee, A. (2007). Using logistic regression to detect bias when multiple groups are tested. South African Journal of Psychology, 37, 47 – 61

[10] Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. Educational and Psychological Measurement, 65:6, 935-953.

[11] Le, V. (1999). Identifying differential item functioning on the NELS: 88 History Achievement Test. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved November 2011 from ProQuest Journals.

[12] Mazor, K. E., Kanjee, A., and Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. Journal of Educational Measurement, 32, 131 – 144.

[13] Navas-Ara, M. J., & Gomez-Benito, J. (2002). Effects of ability scale purification on the identification of differential item functioning. European Journal of Psychological Assessment. Vol. 18, No. 1, pp. 9-15.

[14] Nijenhuis, J. T., Tolboom, E., Resing, W., & Bleichrodt, N. (2004). Does cultural background influence the intellectual performance of children from immigrant groups? The RAKIT Intelligence Test for Immigrant Children. European Journal of Psychological Assessment. Vol. 20, No. 1, pp. 10-26.

[15] Osterlind, Steven J. (1983). Test Item Bias. Sage Publications, Inc., California.

[16] Osterlind, S. J. & Everson, H. T. (2009). Differential Item Functioning. 2nd Edition, CA: Sage Publications.

[17] Penfield, R. P. & Camilli, G. (2007). Differential item functioning and item bias. In C. Rao & S. Sinharay (Eds.) Handbook of Statistics Psychometrics, Vol. 26.

[18] Reynolds R. C., Livingston, R. B., & Willson, V. (2006). Measurement and Assessment in Education. Boston: Pearson.

[19] Rogers, H. J. & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement, 17, 105 – 116.

[20] Roussos, L. A. & Stout, W. (2004). Differential item functioning analysis: Detecting DIF items and testing DIF hypotheses. In D. Kaplan (Ed.) The Sage Handbook of Quantitative Methodology for Social Sciences. Thousand Oaks: Sage.

[21] Sheppard, R., Han, K., Colarell, S. M., Dai, G., & King, D. W. (2006). Differential Item Functioning by Sex and Race in the Hogan Personality Inventory. SAGE Publications. Retrieved June 1, 2010 from ProQuest Educational Journals

[22] Stoneberg, B. D. (2004). A study of gender-based differential item functioning (DIF) in the Spring 2003 Idaho Standards Achievement Tests applying the Simultaneous Bias Test (SIBTEST) and the Mantel-Haenzel Chi Square Test. The University of Maryland Measurement, Statistics, and Evaluation Department and the National Center for Education Statistics (NCES) Assessment Division.

[23] Swaminathan, H. and Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361 – 370.

[24] Wang, N. & Lane, S. (1996). Detection of gender-related differential item functioning in a mathematics performance test. Applied Measurement in Education, 9:2, 175-199.

[25] Wiberg, M. (2009). Differential item functioning in mastery tests: A comparison of three methods using real data. International Journal of Testing. 9, 41-59.

[26] Wolf, L. R. & Phyllis, R. (1990). The SAT gender gap. Women and Languge. Vol. 13, Issue 2. Retrieved September 26, 2011 from ProQuest Journals.

[27] Zheng, Y., Geirl, M. J., & Cui, Y. (2007). Using real data to compare DIF detection and effect size measures among Mantel-Haenzel, SIBTEST, and Logistic Regression procedures. Paper presented at NCME 2007, Chicago.