

Identifying Biased Test Items by Differential Item Functioning Analysis Using Contingency Table Approaches: A Comparative Study

Jose Q. Pedrajita
*College of Education
University of the Philippines, Diliman*

Vivien M. Talisayon
*College of Education
University of the Philippines, Diliman*

Abstract

This study identified biased test items through differential item functioning analysis using four contingency table approaches: Chi-Square, Distracter Response Analysis, Logistic Regression, and Mantel-Haenszel Statistic. The study made use of test scores of 200 junior high school students. One hundred students came from a public school, and the other 100 were private school examinees. One hundred students were males and 100 were females. Basing from their English II grades, 95 students were classified as low ability and 105 as high ability students. A researcher-constructed and validated Chemistry Achievement Test was used as research instrument. The results from the four methods used were compared, and it was found that school type, gender, and English ability bias exists. There was a high degree of agreement between the Logistic Regression and the Mantel-Haenszel Statistic in identifying biased test items.

Keywords: item bias, item bias methods, differential item functioning, measure of bias

Questions of test bias are closely related to questions of test validity. A test is valid if it measures what it purports to measure and invalid if it does not. Bias or systematic error is a kind of invalidity that arises relative to groups. It is typically suspected that there is test bias when a given identifiable group scores low or high on a test relative to some other groups, other things being equal. Bias is a major

factor for tests considered unfair, inconstant, and contaminated by extraneous factors (Camilli and Shepard, 1994).

Biased items may (1) result in differential performance for individuals of the same ability from different ethnic, sex, or cultural groups; (2) lower the average score of a particular group; (3) contain content or language that is differentially familiar to matched groups of examinees; (4) contain sources of difficulty that are irrelevant or extraneous to the construct being tested, thus adversely affecting test performance; (5) ask for information that disadvantaged children/students have not had equal opportunity to learn.

Furthermore, biased items may (1) contain content which may be radically different from a particular subgroup of students' life experiences but the assessment results may be interpreted without taking such differences into proper consideration; (2) be representing single-gender negative stereotypes, rather than a balance of gender accomplishments; (3) contain clues that would facilitate the performance of one group over another; (4) contain inadequacies or ambiguities in the test instructions, item stem, keyed response, or distracters.

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.

One way to investigate bias at the item level is through *differential item functioning* (DIF) analysis. 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 & Shepard, 1994; Kamata & Vaughn, 2004).

Differential item functioning refers to the differing probabilities of success on an item of examinees of the same ability but belonging to different groups. 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. 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 & Shepard, 1994).

To date, however, there has been a dearth of studies on item bias and comparison of item bias methods conducted in the Philippines. There are no empirical studies done of test bias and comparison of item bias methods, test users are left with very little certainty about the validity and cultural appropriateness of the measures they take. It is thus essential to raise the consciousness of assessment

practitioners regarding how unacceptable it is to use tests before having undertaken bias studies, particularly for high stake examinations.

The comparison of item bias methods is an important practical concern, since both the size of sample required and the cost associated with the procedures differ widely. If all the bias approaches were to identify the same items as biased, one could use the simplest and least expensive approach. However, if the approaches identify different items as biased, it becomes necessary to determine those methods which are most valid (Osterlind, 1983).

There is a need to empirically compare the various methods, specifically, the contingency table (CT) approaches. This could help fill the knowledge gap and may lead to a better understanding of the usefulness of the said approaches. The present study represents an attempt to meet this need.

This study looked into biased test items between public and private, male and female, and low and high English ability examinees in a researcher-constructed and validated Chemistry Achievement Test through *differential item functioning* analysis. It also looked into the agreement among the DIF approaches in identifying biased test items.

Methodology

This study employed the descriptive-comparative research design. Three reference/focal group combinations 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. For each pair of matched group the total number of examinees adds up to 200, which was the total sample in this study. Each pair of matched examinees were matched by section and total score.

The *Statistical Analysis System* (SAS) computer software was used in the analysis of data. The analysis of data involved (a) assignment of examinees' test papers to the comparison group 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) detecting and identifying biased test items between the comparison group.

The *chi-square method* (X^2) examines the likelihood or probability of test takers from different groups with the same ability levels correctly responding to an item. The *chi-square method* involved the following steps: (1) establishing the ability levels on the total score scale; (2) placing the data in contingency tables; and (3) significance testing. 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 focal groups.

The *distracter response analysis* (DRA) examines the incorrect alternatives to a test item for differences in patterns of response among different subgroups of a population. In the *distracter response analysis* the steps were: (1) preparation of a matrix of choice-response alternatives for the test items under consideration; (2) placing the data in a series of 2 x 2 contingency tables; and (3) hypothesis testing. The hypothesis under test is that *there is no significant difference in proportions selecting distracters on the test items between the reference and focal groups.*

The *logistic regression* (LR) is a kind of regression analysis often used when the dependent variable is dichotomous and scored 0 or 1. It can also be used when the dependent variable has more than two categories. It is usually used for predicting whether something will happen or not – anything that can be expressed as Event/Non-Event. 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.*

The *mantel-haenszel statistic* (MH) is a non-parametric contingency table procedure commonly used to perform statistical test for uniform DIF. When the magnitude of DIF is the same across all ability levels, it is referred to as *uniform DIF*. On the other hand, it is referred to as *non-uniform DIF*, when the magnitude of DIF is not consistent across ability levels. MH yields a chi-square test with one degree of freedom to test the null hypothesis that *there is no relation between group membership and test performance on one item after controlling for total test score.* Aside from the statistical significance of the obtained chi square value, the MH procedure is also used to estimate a ratio, the log odds ratio (β_{MH}), which yields a measure of effect size for evaluating the amount of DIF that is present. This ratio value was rescaled as $D = -2.35\beta_{MH}$ to produce the delta-MH (D-MH). A positive D-MH indicates DIF in favor of a focal group, and a negative value signifies DIF in favor of a reference group. The degrees of DIF in test items are labeled A, B, and C (ETS category) to indicate **negligible**, **moderate**, and **large** amounts of DIF (Gierl, 1999).

These categories are defined as follows: a) A items have D-MH values which do not significantly differ from 0 and smaller than 1.0 in absolute value; b) B items have D-MH values which significantly differ from zero and either D-MH not significantly greater than 1.0 and or D-MH smaller than 1.5 in absolute value; and c) C items have D-MH values which is both significantly greater than 1.0 and greater than 1.5 in absolute value. Items with A- and B-level statistical ratings were considered unbiased, while, items falling into category C were inferred to have large DIF and therefore biased.

The *logistic regression* and the *mantel-haenszel statistic* involved the following steps: *First*, test data were coded and scored. For each examinee it must have (a) a code or label for group membership, (b) the actual response (right or wrong) for each item, and (c) total score on the test. *Second*, to prepare for a DIF analysis, data for any given item were organized into a tabular form that is commonly referred to as a three-way contingency table. *Third*, was the statistical analysis for detecting and testing for differential item functioning and item bias.

In this paper, all tests of hypotheses were carried out at the 0.05 alpha levels. Table 1 shows the statistical criteria for identifying biased test items.

Table 1
Statistical Criteria for Identifying Biased Items

DIF Approaches	Focus of Analysis	Measure of Bias
<i>Chi-Square</i>	<i>Difference in proportions attaining a correct response across score categorie</i>	<i>Significance of chi-square</i>
<i>Distracter Response Analysis</i>	<i>Difference in proportions selecting distracters</i>	<i>Significance of chi-square</i>
<i>Logistic Regression</i>	<i>Odds of getting the item right</i>	<i>Significance of chi—square</i>
<i>Mantel-Haenszel Statistic</i>	<i>Performing statistical test for DIF effect</i>	<i>Significance of ch-square and large DIF effect</i>

The common measure of bias is the significance of the obtained chi square value. A significant chi square value indicates: (1) *difference in proportion attaining a correct response across total score categories* for the X^2 procedure; (2) *difference in proportions selecting distracters* for the DRA; (3) *difference in the odds of getting an item right between the reference/focal groups compared* for the LR; and (4) *large DIF effect* for the MH Statistic. The agreement between and among two, three or all of the methods is indicated by their obtained measure of bias. If any two, three or all of the four methods similarly obtained a statistically significant measure of bias (chi square value) on an item or groups of items, such methods were in agreement. If not, there is disagreement.

School type, gender, and ability bias refers to the differing probabilities of success on an item between the public and private, the male and female, and the low and high ability examinees, respectively.

Figure 1 shows the methodological flowchart of the study.

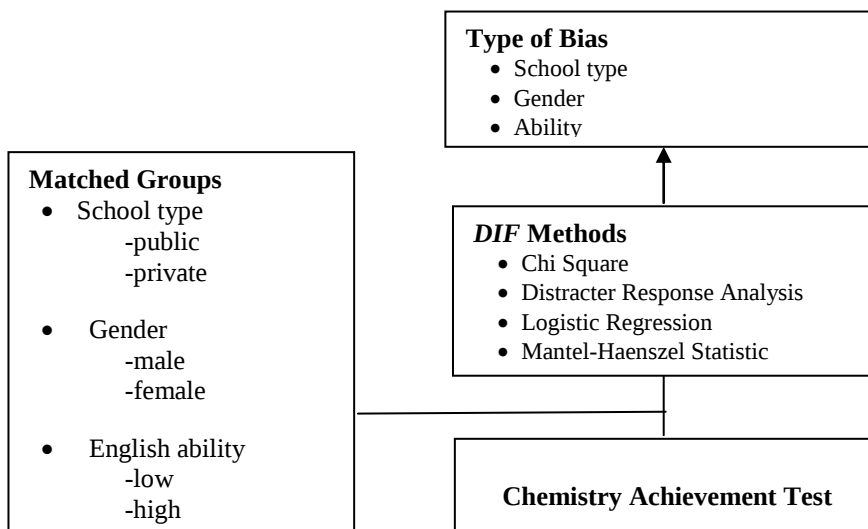


Figure 1
Methodological Flowchart of the Study

It shows that the original chemistry achievement test was administered to the matched groups of examinees. Thereafter, the examinees' scores were subjected to each of the *DIF* methods to identify items indicating school type, gender and ability bias. *School type, gender, and ability bias* was determined from the analysis of the public/private, male/female, and low/high English ability examinees, respectively.

Results

Differential item functioning analysis

Table 2 shows the biased items identified in the DIF analysis between the public and the private school examinees.

Table 2
Biased Items Detected in the Public/Private School Matched Examinees

Items	Concept/Skills Measured	X ²	DRA Biased Against	LR	MH
1	gas property illustrated by garbage smell entering the house	Pvt*	Pvt*	Pvt*	Pvt*
2	element with Latin name “aurum”			Pub*	Pub*
3	chemical bond which held together two atoms in a molecule by the transfer of an electron from one atom to the other	Pvt*	Pvt*	Pvt*	Pvt*
5	Filipino scientist who pioneered in the use of biogas/biomass as a source of energy	Pvt*	Pvt*	Pvt*	
8	definition of valence electrons		Pub*	Pub*	Pub*
9	description of dialysis	Pvt*	Pvt*	Pvt*	Pvt*
10	volume of a cube			Pvt*	Pvt*
13	new pressure of the gas when the volume is compressed to a smaller quantity	Pub*	Pub*	Pub*	Pub*
14	problem on Boyle’s Law			Pub*	Pub*
16	how the chemical and molecular formula of sodium sulfate is correctly written	Pub*	Pub*	Pub*	Pub*
19	solving for the molar mass of Fe ₂ O ₃	Pvt*	Pvt*	Pvt*	Pvt*
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		Pvt*	Pvt*	Pvt*
22	volume conversion	Pub*	Pub*	Pub*	Pub*
26	indicators of chemical change			Pvt*	Pvt*
30	correct position of Chlorine in the periodic table	Pvt*	Pvt*	Pvt*	Pvt*
31	indicator of a balanced chemical equation	Pvt*			
32	which chemical equation is balanced			Pub*	Pub*

(Table 2 Cont.)

33	identify the reactants in the given chemical equation	Pvt*	Pvt*	Pvt*	Pvt*
35	identify which principle is true of different substances having an equal number of moles		Pvt*		
36	classification of a solution which changes red litmus paper to blue		Pub*	Pub*	Pub*
37	factors which increases the solubility of a solute	Pub*		Pub*	Pub*
40	evidences of chemical change			Pub*	Pub*
41	laws which govern changes in matter		Pub*	Pub*	Pub*
43	properties of gases		Pvt*		
46	components of a solution				Pub*
47	strategy which is most probable in proving the given hypothesis in the given experiment	Pvt*	Pvt*	Pvt*	Pvt*
50	factor which causes the nails to rust		Pvt*		

*p < 0.05 Pub = Public Pvt = Private

X² Analysis. The chi-square analysis identified 13 biased items between the public and private school examinees. Nine of which, items 1, 3, 5, 9, 19, 30, 31, 33, and 47, were biased against the private school examinees. That is, the *probability of success* on these items favored the public school examinees. Whereas, four items, items 13, 16, 22, and 37, were biased against the public school examinees, indicating that in each of these items, the *probability of success* favored the private school examinees. Thus, the null hypothesis that *there is no significant difference in proportions attaining a correct response across total score categories on the test items between the public and private school examinees* is rejected in favor of the alternative hypothesis.

DRA Analysis. The distracter response analysis revealed 18 items which indicate bias 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 biased against the private school examinees. In each of these items, a large number of private school examinees was attracted to the incorrect options indicating

unfamiliarity with the concept reflected in the items. Hence, the *probability of success* on these test items favored the public school examinees. Whereas, six, items, 8, 13, 16, 22, 36, and 41, were biased against the public school examinees. In these items, most of the public school examinees were attracted to the incorrect options, indicating less familiarity with the concept reflected in these items. That is, the *probability of success* on these test items favored the private school examinees. Thus, the null hypothesis that *there is no significant difference in proportions selecting distracters on the test items between the public and private school examinees* is rejected in favor of the alternative hypothesis.

LR Analysis. The LR analysis identified 22 items which indicate bias 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 the twenty-two biased items, 1, 3, 5, 9, 10, 19, 21, 26, 30, 33, and 47 were 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 biased against the public school examinees. In each of these items, the *odds of getting an item right* favored the private school examinees. Hence, the null hypothesis that *the population value is zero for either the difference between the proportions correct or the log odds ratio on the test items between the public and private school examinees* is rejected in favor of the alternative hypothesis.

MH Analysis. The MH analysis between the public and the private school examinees showed that majority of the multiple choice items did not display differential item functioning. Using the three-tiered ratings, 19 of the 50 items displayed negligible effects (A items); 9 of the 50 items displayed moderate effects (B items); and 22 of the 50 items displayed large effects (C items).

Of the 22 C 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 C items obtained a significant MH chi square value and a negative delta-MH greater than 1.5 in magnitude, signifying DIF in favor of the public school examinees. Whereas, twelve items, items 2, 8, 13, 14, 16, 22, 32, 36, 37, 40, 41, and 46 favored the private school examinees. Each of these twelve C items obtained a significant MH chi square value and a positive delta-MH greater than 1.5 in magnitude, indicative of DIF in favor of the private school examinees. Thus, the null hypothesis that *there is no significant relationship between group membership and test performance on the test items between the public and private school examinees* is rejected in favor of the alternative hypothesis.

Table 3 shows the biased items detected in the DIF analysis between the male and the female examinees.

Table 3
Biased Items Detected in the Male/Female Matched Examinees

Items	Concept/Skills Measured	X ²	DRA Biased Against	LR	MH
1	gas property illustrated by garbage smell entering the house			M*	M*
3	chemical bond which held together two atoms in a molecule by the transfer of an electron from one atom to the other			M*	M*
17	electron configuration of the element Sodium	F*		F*	F*
27	options which illustrate the compressibility of gases		F*	F*	F*
34	definition of reaction reversibility			F*	F*
42	principles of Kinetic Molecular Theory		M*	M*	M*
47	strategy which is most probable in proving the given hypothesis in the given experiment			M*	M*
*p < .05 M = Male F = Female					

X² Analysis. The chi-square analysis reveals that only one item, item 17, was found biased between the male and the female examinees. The matched groups had different probability of success on the item. That is, the *probability of success* on this item favored the male examinees. Thus, the null hypothesis that *there is no significant difference in proportions attaining a correct response across total score categories on the test items between the male and female examinees* is rejected in favor of the alternative hypothesis.

DRA Analysis. The DRA analysis showed 2 items which indicate bias between the male and the female examinees. They were items 27 and 42. Of the two, one was biased against the female examinees and the other was biased against the male examinees.

Item 27 was biased against the female examinees. The female examinees obtained a large number of responses in the incorrect options, indicating less familiarity with the concept reflected in the item. These incorrect options were seemingly plausible for the said examinees. Hence, the *probability of success* on this test item favored the male examinees. Conversely, item 42 was biased against the male examinees. The male examinees obtained a differentially large number of

responses in the incorrect options. These incorrect options seemed likely to be the correct answer on their part. Hence, the *probability of success* on this test item favored the female examinees. Thus, the null hypothesis that *there is no significant difference in proportions selecting distracters on the test items between the male and female examinees* is rejected in favor of the alternative hypothesis.

LR Analysis. 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 biased against the female examinees. In these items, the *odds of getting an item right* favored the male examinees. Whereas, four items, 1, 3, 42, and 47, were biased against the female examinees. In these items, the *odds of getting an item right* favored the female examinees. Thus, the null hypothesis that *the population value is zero for either the difference between the proportions correct or the log odds ratio on the test items between the male and female examinees* is rejected in favor of the alternative hypothesis.

MH Analysis. The MH analysis between the male and the female examinees revealed that majority of the multiple choice items did not display differential item functioning. Using the three-tiered ratings, 37 of the 50 items displayed negligible effects (A items), 6 of the 50 items displayed moderate effects (B items), and 7 of the 50 items displayed large effects (C items).

Of the seven C items, three favored the male examinees. These were items 17, 27, and 34. Each of the three C items obtained a significant MH chi square value and a negative delta-MH greater than 1.5 in magnitude, signifying DIF in favor of the male examinees. Whereas, four items, 1, 3, 42, and 47, favored the female examinees. These four C items obtained a significant MH chi square value and a positive delta-MH greater than 1.5 in magnitude, indicative of DIF in favor of the female examinees. Thus, the null hypothesis that *there is no significant relationship between group membership and test performance on the test items between the male and the female examinees* is rejected in favor of the alternative hypothesis.

This finding is apparently similar to Gierl's (1999) study which 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.

Table 4 shows the biased items detected in the DIF analysis between the low and the high ability examinees.

Table 4
Biased Items Detected in the Low/High Ability Matched Examinees

Items	Concept/Skills Measured	X ²	DRA Biased Against	LR	MH
2	element with Latin name “aurum”	L*		L*	L*
3	chemical bond which held together two atoms in a molecule by the transfer of an electron from one atom to the other	L*	L*	L*	L*
6	scope of chemistry	L*	L*	L*	L*
7	property of gases that best describes the foul odor of a nearby garbage dump		L*		
8	correct definition of valence electrons	L*	L*	L*	L*
13	new pressure of the gas when the volume is compressed to a smaller quantity	L*	L*	L*	L*
15	problem-solving on Charles’ Law		L*		
17	electron configuration of the element Sodium				L*
19	solving for the molar mass of Fe ₂ O ₂	L*	L*	L*	L*
22	volume conversion		L*	L*	L*
29	valence electrons of the Chlorine atoms			H*	H*
30	correct position of Chlorine in the periodic table	L*	L*		
36	classification of a solution which changes red litmus paper to blue			L*	L*
38	in which solution water is a solute			H*	H*
45	in which situation the process of oxidation is common			H*	H*
48	correct formula in solving for the new volume of the gas		L*	L*	L*
50	factor which causes the nails to rust		L*	L*	L*
*p < .05	L = Low Ability	H = High Ability			

X² Analysis. The chi-square analysis between the low and high ability examinees identified seven biased items, namely items 2, 3, 6, 8, 13, 19, and 30. All of these items were biased against the low ability examinees. That is, the *probability of success* in these items favored the high ability examinees. None of the items, however, were biased against the high ability examinees. Nevertheless, the null hypothesis that *there is no significant difference in proportions attaining a correct response across total score categories on the test items between the low and the high ability examinees* is rejected in favor of the alternative hypothesis.

DRA Analysis. The distracter response analysis 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 biased against the low ability examinees.

In each of these items, one, two or all of the three incorrect options had obtained a large number of responses from the low ability examinees. These incorrect options were seemingly plausible for the said examinees. Hence, the probability of success on these test items favored the high ability examinees. Thus, the null hypothesis that *there is no significant difference in proportions selecting distracters on the test items between the low and the high ability examinees* is rejected in favor of the alternative hypothesis.

LR Analysis. The LR analysis discovered 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 29, 38, and 45, were biased against the high ability examinees. In these items, the *odds of getting an item right* favored the low ability examinees. Whereas, ten items, 2, 3, 6, 8, 13, 19, 22, 36, 48, and 50, were biased against the low ability examinees. In these items, the *odds of getting an item right* favored the high ability examinees. Thus, the null hypothesis that *the population value is zero for either the difference between the proportions correct or the log odds ratio on the test items between the low and the high ability examinees* is rejected in favor of the alternative hypothesis.

MH Analysis. In the MH analysis between the low and the high ability examinees, majority of the multiple choice items did not display differential item functioning. Using the three-tiered ratings, 28 of the 50 items displayed negligible effects (A items); 8 of the 50 items displayed moderate effects (B items); and 14 of the 50 items displayed large effects (C items).

Of the fourteen C items, three favored the low ability examinees. They were items 29, 38, and 45. These three C items obtained a significant MH chi square value and a negative delta-MH greater than 1.5 in magnitude, signifying DIF in favor of the low ability examinees. Whereas, eleven items, items 2, 3, 6, 8, 13, 17, 19, 22, 36, 48, and 50, favored the high ability examinees. These C items obtained a

significant MH chi square value and a positive delta-MH greater than 1.5 in magnitude, indicative of DIF in favor of the high ability group. Thus, the null hypothesis that *there is no significant relationship between group membership and test performance on the test items between the low and the high ability examinees* is rejected in favor of the alternative hypothesis.

Agreement of the DIF methods on biased items detected

Table 5 shows the agreement between and among the DIF methods in detecting biased items. The upper column contains the biased items against the *private, female and high ability* examinees, while, the lower column contains the biased items against the *public, male, and low ability* examinees.

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 methods.

Ten items were identically identified by the four methods. Seven of which, items 1, 3, 9, 19, 30, 33, and 47, were biased against the private school examinees. These items have indices of difficulty within .5 to .78. That is, these difficulty indices indicate that these were relatively easy items, being above the .5 level of difficulty. However, item 1 was also commonly identified in the LR and MH analyses as biased against the male examinees. Item 3 was also identically identified by the four methods as biased against the low ability examinees and further identified in both the LR and MH analyses as biased against the male examinees. Moreover, item 19 was also identically identified by the four methods as biased against the low ability examinees. Item 30 was also identified as biased 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 biased against the male examinees. Whereas, three, items 13, 16, and 22 were biased against the public school examinees. They have indices of difficulty which ranged from .16 to .36. These difficulty indices indicate that these items are relatively difficult items, being lower than the .5 level of difficulty. Thus, the relatively easy items were biased against the private school examinees and the relatively difficult items were biased against the public school examinees. However, item 13 was also identically identified by the four methods as biased against the low ability examinees. Moreover, item 22 was likewise commonly identified in the DRA, LR, and MH analysis as biased against the low ability examinees.

Table 5
Agreement of the DIF Methods on Biased Items Detected

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

Four items were commonly identified by the DRA, LR, and MH statistic. One of which, item 21, was biased against the private school examinees. Its difficulty index of .78 indicates that it was an easy item. Conversely, items 8, 36, and 41 were biased against the public school examinees. Their difficulty indices ranged from .28 to .54 which means that these items were relatively difficult though they were within the middle range or optimum difficulty level. Moreover, item 8 was commonly identified by the four methods as biased against the low ability examinees. In addition, item 36 was also identified by both LR and MH Statistic as biased against the low ability examinees.

Only one item, item 5, was identically identified as biased against the private school examinees in the X^2 , DRA, and LR analyses. It has difficulty index of .76, indicating that it is an easy item.

Another lone item, item 37, was commonly identified in the X^2 , LR, and MH analyses as biased against the public school examinees. It has difficulty index of .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 biased against the private school examinees. They have difficulty index of .7 and .84, respectively, indicating that these were relatively easy items. Whereas, four items, items 2, 14, 32, and 40 were biased against the public school examinees. Their difficulty indices were .21, .34, .46, and .79 respectively. These difficulty indices indicate that these items were relatively difficult, with the exception of item 2. However, item 2 was also commonly identified as biased 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 biased against the private school examinees. Their difficulty indices were .33, .64, and .66, respectively. Though all of them belong to the middle range of difficulty, items 43 and 50 were relatively easier than item 35. Moreover, item 50 was also commonly identified in the DRA, LR, and MH analyses as biased against the low ability examinees.

A lone item, item 31, was singly identified only in the X^2 analysis as biased against the private school examinees. It has difficulty index of .64 indicating that it was a relatively easy item, being above the .5 level of difficulty.

Another single item, item 46, was identified only by the MH Statistic as biased against the public school examinees. It has difficulty index of .73 indicating that it is a relatively easy item, being above the .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 biased items against the private school examinees were relatively easier items, mostly within the middle and upper ranges of difficulty levels. Whereas, 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 methods yielded very similar results. Both identified 22 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 singly or commonly identified by one, two, three, or all of the four DIF methods.

Item 17 was commonly identified in the X^2 , LR, and MH analyses as biased against the female examinees. It has difficulty index of .86, indicating that it is a very easy item. Moreover, item 17 was also identified as biased 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 biased item.

Items 27 and 42 were commonly identified in the DRA, LR, and MH analyses. Item 27 was biased against the female examinees. It has difficulty index of .58, indicating that it is a relatively easier item. On the other hand, item 42 was biased against the male examinees. It has difficulty index of .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 biased against the female examinees. Its difficulty index is .18, indicating that it is a difficult item. Conversely, item 47 was biased against the male examinees. Its difficulty index was .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 methods. 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 solely or commonly identified by one, two, three, or all of the four DIF methods.

Item 6 was commonly identified by the four methods as biased against the low ability examinees. Item 6 has difficulty index of .79, indicating that it is a relatively easy item. Item 48 was also commonly identified in the DRA, LR, and MH analysis as biased against the low ability examinees. Its difficulty index of .7 indicates that it is an easier item.

Items 29, 38, and 45 were commonly identified as biased against the high ability examinees by the LR and the MH Statistic. Their difficulty indices were .38, .41, and .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 .74 and .56, respectively, indicating relatively easier items because, though in the middle range of difficulty ranges, their difficulty indices were above the .5 difficulty index.

A closer scrutiny of the 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 .5 index of difficulty, except items 13, 22, and 36. On the other hand, the biased items against the high ability examinees have difficulty indices within the middle range or optimum difficulty level and less than the .5 level of difficulty. Hence, the pattern of bias was largely toward the low ability examinees.

The LR and the MH analysis 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 item more than the LR.

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

Discussion

The DIF analysis between the public and the private school examinees detected thirteen biased items in the X^2 analysis, nine of which were biased against the private school examinees, while, four items were biased against the public school examinees; eighteen items in the DRA analysis, twelve of which were biased against the private school examinees, while six items were biased against the public school examinees; twenty-two items in the LR analysis, eleven of which were biased against the private school examinees, while the other eleven were biased against the public school examinees; and also twenty-two items in the MH analysis, ten of which were biased against the private school examinees, while, twelve items were biased against the public school examinees.

The analysis showed that there were more biased items against the private than against the public school examinees. This indicates the test was relatively more difficult for the private than for the public school examinees. The affected group was less likely familiar with the concept reflected in the biased items. Moreover, the disadvantaged group could have received inferior lesson or may have been denied of learning experience necessary to obtain a correct response on the biased items.

The DIF analysis between the male and the female examinees identified only one item in the X^2 analysis which was biased against the female examinees; two items in the DRA, one was biased against the female examinees and the other one was biased against the male examinees; seven items in the LR analysis, three of which were biased against the female examinees, while four were biased against the male examinees; and seven items (C items) in the MH analysis, three of which were biased against the female examinees, while four were biased against the male examinees.

The analysis indicated that the biased items were relatively more difficult for the disadvantaged group. The concepts, information and/or skill reflected in the content of the biased items were less likely familiar to the affected group. There may have been less opportunity on the part of the affected examinees to become acquainted with the concepts and principles involved in the biased items. Apparently, the two groups had not had equal opportunity for learning experience related to the content of the biased items. The question may be poorly worded and unfamiliar to the affected group. Moreover, the affected group could have received limited lesson or may have been denied of learning experience necessary in obtaining a correct response on the biased items. Overall, the results indicate that the Chemistry Achievement Test is generally fair between the male and the female examinees.

The DIF analysis between the low and the high ability examinees identified seven biased items in the X^2 analysis, all of which were biased against the low ability examinees; eleven in the Distracter analysis; thirteen in the LR analysis, three of which were biased against the high ability examinees and ten were biased against the low ability examinees; and fourteen in the MH analysis, eleven of which were biased against the low ability examinees and three were biased against the high ability examinees.

The analysis showed that the test was relatively difficult for the low ability examinees. There were more biased items against the low ability examinees than against the high ability examinees. The low ability examinees were less likely acquainted with the concept reflected in these items. The performance differences in the low/high ability comparison groups may have been due to differences in ability in interpreting or comprehending written English. Language ability is relevant to the purpose of testing. It is an important dimension in any test which requires more than the most minimal reading. Moreover, the low ability examinees could have received inferior lessons or may have been deprived of learning experience related to such items.

A closer scrutiny of the 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. That is, these items have difficulty indices ranging from optimum difficulty to very easy, mostly higher than the .5 index of difficulty, except items 13, 22, and 36. On the other hand, the biased items against the high ability

examinees have difficulty indices only within the middle range or optimum difficulty level and lesser than the .5 level of difficulty. Hence, the pattern of bias was largely against the low ability examinees.

The LR and the MH Statistic yielded very similar results. Each identified 13 identical biased items, except for the MH chi square which identified item 17, giving it an extra item more than the LR.

The findings in the three matched group analysis deserve further comment. *First*, the number of items exhibiting bias with both the LR and the MH procedures seems high. Apparently, both LR and MH are the most sensitive among the four item bias techniques. *Second*, consistent with earlier research, regardless of which criterion the comparison is based on, the MH and the LR procedures result in similar number of items (and similar items) being identified (Rogers & Swaminathan, 1993). Thus, there is a high degree of correspondence between the LR and the MH procedures when either one or two ability estimates are included in the analysis. LR has shown that under comparable conditions, when matching is based on a single test score, it produces results that are extremely similar to those produced using the MH Statistic (Swaminathan, 1990; Mazor et al., 1995).

The four methods for detecting item bias may be evaluated not only in terms of logical appeal or statistical adequacy, but in terms of external evidence of validity. Some possible types of validity evidence for a bias technique would be a demonstration that: (1) the procedure is not selecting items at random; and (2) the results obtained with different methods tend to agree. Perfect agreement would probably not be expected, due to differences in the assumptions and limitations of the various methods. Thus, the LR and MH procedures appear to have demonstrated the external validity evidence mentioned above. Hence, these two approaches are widely implemented in DIF detections.

The presence of *school type bias*, *gender bias*, and *ability bias* in the Chemistry Achievement Test can be attributed to the: (1) discrepancies in the curriculum of the public and private school; (2) less familiarity with the content of the biased items which caused the examinees to be attracted to the incorrect options; (3) ambiguities in the item stem, keyed response, or distracter; (4) disparity in the matched examinees' exposure to the information, concepts or skills reflected in the biased items; (5) quality of teaching and lessons received by the examinees; and/or (6) inability of the matched examinees to comprehend or understand the concepts reflected on the biased items.

Conclusions

The results of the differential item functioning analysis showed that there were statistically biased test items between (1) the public and the private school examinees, (2) the male and the female examinees, and (3) the low and the high ability examinees. Hence, *school type bias*, *gender bias*, and *ability bias* were

present in the Chemistry Achievement Test. Overall, it appears that students from public schools performed better than those from private schools; male and female examinees performed fairly; and low ability examinees performed miserably than the high English ability examinees in the Chemistry Achievement Test. Ability bias was heavily tilted against the low ability examinees.

There were agreements between and among the item bias methods in the identity and number of biased items detected. 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.

Investigating bias at the item level is particularly useful in the process of test development, in which biased items are revised or removed. This is a legitimate and important process in an attempt to achieve test equity (Kamata & Vaughn, 2004). Test equity is primarily achieved by ensuring that a test measures only construct-relevant differences between subpopulations (Messick, 1989 as cited in Kamata & Vaughn, 2004). If test equity is not achieved, a test or test item is biased toward a particular subpopulation of the test taking population. Statistically, a test or test item is said to be biased if the expected test or item scores are not the same for examinees from different subpopulations, given the same level of trait that the test intends to measure (Kamata & Vaughn, 2004).

In deciding which item bias method to use, it is appropriate to choose methods which are most valid. Valid methods may be very sensitive and may have a very high detection rate in identifying biased test items. But it is better for test development for it could identify all items which are possibly biased, and then to eliminate or revise 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 not be identified and may remain part of the test content, thereby, still affecting and contaminating the validity and reliability of the test.

Item bias methods with high detection rate are preferable over those with low detection rate in purifying assessment instrument. That is, test items should be free of bias.

Findings can significantly contribute to educational measurement. The use of statistical methods in identifying biased test items is a relatively better kind of item analysis. By subjecting test items to item bias detection approaches, test items which were unfairly difficult and widely discriminating for a particular group of examinees are determined. By eliminating, replacing, or revising these biased items a valid, reliable, and fairer test would be made.

Recommendations

Test experts and developers should use contingency table (CT) methods, particularly the LR and MH methods, in item bias detection. These two methods are viable in the detection of DIF and are widely implemented in both test construction and research settings.

Educational evaluation practitioners who are engaged in item bias detection should use Logistic Regression or Mantel-Haenszel Statistic for bias correction. That is, identified biased items should be revised or replaced. Then re-administer the test and subject it anew to item bias detection in order to further refine and purify the required item content of a test. This process could make differentially functioning items between groups of interest be more valid, reliable, and fair. Bias correction can maintain or improve the measurement qualities of a test such as its content validity, concurrent validity, and internal consistency reliability.

It is also recommended that a study be conducted using Logistic Regression and/or Mantel-Haenszel Statistic 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.

Future studies should focus on other psychometric issues not addressed in this study. These include matters related to comparative study of *Item Response Theory* (IRT) and *Contingency Table* (CT) methods on any relevant psychometric issue, such as test equating and item banking.

There is a need for bias testing especially for very important tests like entrance examinations and professional licensure examinations.

It is also recommended that further studies be conducted to go beyond detecting biased items and obtain additional information about biased items. Some items may show larger magnitude of bias, while some others show relatively small magnitude of bias. In such a situation, it is of interest to investigate sources of such variation.

References

- Camilli, G. & Shepard, L. (1994). *Methods for identifying biased test items*. Vol. 4. California: Sage Publications.
- Gierl, M. J. (1999). *Differential item functioning on the Alberta education social studies 30 diploma examination*. Canadian Social Studies, 33(2).
- Kamata, A. & Vaughn B. (2004). *An introduction to differential item functioning analysis*. Learning Disabilities: A Contemporary Journal 2(2), 49-69.

- Kanjee, A. (2007). *Using logistic regression to detect bias when multiple groups are tested*. South African Journal of Psychology, 37, 47-61.
- Mazor, K. E., Kanjee, A., & 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.
- Osterlind, Steven J. (1983). *Test item bias*. California: Sage Publications.
- 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.
- Swaminathan, H. & Rogers, H. J. (1990). *Detecting differential item functioning using logistic regression procedures*. Journal of Educational Measurement, 27, 361-370.

Date received: June 2008

Jose Q. Pedrajita is a full time faculty member of the Research and Evaluation Area of the College of Education, University of the Philippines, Diliman.

Vivien M. Talisayon is dean of and professor in the College of Education, University of the Philippines, Diliman.