World Journal of Educational Research
ISSN 2375-9771 (Print) ISSN 2333-5998 (Online)
Vol. 4, No. 1, 2017
www.scholink.org/ojs/index.php/wjer

Correction of Differentially Functioning Items: Basis for Maintaining and Enhancing Test Validity and Reliability

Jose Q. Pedrajita[1*]

[1] University of the Philippines, Diliman, Quezon City, Philippines

[*] Jose Q. Pedrajita, E-mail: josepedrajita@gmail.com

Received: November 6, 2016 Accepted: November 26, 2016 Online Published: December 9, 2016

doi:10.22158/wjer.v4n1p62 URL: http://dx.doi.org/10.22158/wjer.v4n1p62

Abstract

This study looked into differentially functioning items in a Chemistry Achievement Test. It also examined the effect of eliminating differentially functioning items on the content and concurrent validity, and internal consistency reliability of the test. Test scores of two hundred junior high school students matched on school type were subjected to Differential Item Functioning (DIF) analysis. One hundred students came from a public school, while the other 100 were private school examinees. The descriptive-comparative research design utilizing differential item functioning analysis and validity and reliability analysis was employed. The Chi-Square, Distractor Response Analysis, Logistic Regression, and the Mantel-Haenszel Statistic were the methods used in the DIF analysis. A six-point scale ranging from inadequate to adequate was used to assess the content validity of the test. Pearson r was used in the concurrent validity analysis. The KR-20 formula was used for estimating the internal consistency reliability of the test. The findings revealed the presence of differentially functioning items between the public and private school examinees. The DIF methods differed in the number of differentially functioning items identified. However, there was a high degree of correspondence between the Logistic Regression and Mantel-Haenszel Statistic. After the elimination of the differentially functioning items, the content and the concurrent validity, and the internal consistency reliability differed per DIF method used. The content validity of the test differed ranging from slightly adequate to moderately adequate in the number of items retained. The concurrent validity of the test also differed but all were positive and indicate moderate relationship between the examinees' test scores and their GPA in Science III. Likewise, the internal consistency reliability of the test differed. The more differentially functioning items eliminated, the lesser was the content and concurrent validity, and internal consistency reliability of the test becomes. Elimination of differentially functioning items diminishes content and concurrent validity, and internal consistency reliability, but could be use as basis in enhancing content, concurrent as well as internal consistency reliability by replacing eliminated DIF items.

Keywords

differential item functioning, DIF analysis, internal consistency reliability, content validity, concurrent validity

1. Introduction

A critical step in the development of educational assessment instruments is to ensure that no individual or group responding to the instrument is disadvantaged in any way. This is an important process to achieve test equity. Test equity is primarily achieved by ensuring that a test measures only construct-relevant differences between subpopulation of examinees. If test equity is not achieved, a test or test item is biased toward a particular subpopulation of examinees (Kanjee, 2007).

Test items are subjected to DIF detection techniques to determine whether or not they conform to a given set of psychometric rules in the same way for all persons in a population, regardless of any subgroup membership within that population.

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 is said to be present in a test item when, despite controls for overall test performance, examinees from different groups have a different probability or likelihood 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). 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).

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

According to O'Neill and McPeek (1993), "The fundamental principle of DIF is simple: Examinees who know the same amount about a topic should perform equally well on an item testing that topic regardless of their sex, race or ethnicity".

In this study, four contingency table methods, the Chi-Square, Distractor Response Analysis, Logistic Regression (LR) and the Mantel-Haenszel (MH) Statistic were employed in detecting DIF on the public/private school matched examinees. Based on the result of the analysis in the public/private matched group, it also tested the effect of eliminating these DIF items on the content validity, concurrent validity, and internal consistency reliability of the test (labeled as test versions) after subjecting it to each of the DIF method and compared it to the original item pool of the Chemistry

Achievement Test.

In this study, these four methods for detecting DIF were evaluated further in terms of external evidence of validity. The types of validity evidence for a DIF technique would be a demonstration that: (a) the procedure is not selecting item at random; and (b) 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.

2. Method

This study employed the descriptive-comparative research design. The research instrument was a Multiple Choice Chemistry Achievement Test. It was administered to selected public and private school matched examinees. Thereafter, the examinees' scores were subjected to the four DIF methods to identify items indicating DIF. Differential item functioning refers to the differing probabilities of success on an item(s) between the 100 public and the 100 private school examinees. These examinees were third year high school students taken from the top, middle, and lower class sections of a public and a private school in the Division of City Schools, Quezon City, Philippines.

The preparation of the Chemistry Achievement Test items involved the following steps: (1) development of a Table of Specifications; (2) consultation with 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 following indices of item difficulty and item discrimination were used in deciding whether to discard or retain an item after the item analysis.

Index of Difficulty	Index of Discrimination
91% & above—very easy, to be discarded	.40 and up—very good item
76-90—easy, needs revision	.30-.39—good item
26-75—highly acceptable, optimum difficulty	.20-.29—marginal item
11-25—difficult, needs revision	.19 & below—poor item
10% and below—very difficult, to be discarded	

Items with difficulty indices within .20 to .80 and discrimination indices of .30 to .80 were retained. This means that items with difficulty level of .20 and below (very difficult) and .81 and above (very easy) were discarded. In like manner, items with discrimination indices of .20 to .29 (marginal item) and .19 and below (poor item) were rejected. However, after the item analysis of the 75-item pool, 14 marginal items and 4 poor items were considered for inclusion in the final form to complete the required number of items to 50 in the research instrument. The basis of consideration is that their difficulty levels ranged from easy, optimum difficulty, to difficult which are acceptable difficulty ranges. The items considered for inclusion have discrimination indices of .20 to .29 (marginal items) and .19 and below (poor items). Their difficulty indices ranged from .20 to .80 which is the acceptable range of difficulty for test items. The poor items were items 18, 49, 63, and 72 with discrimination indices ranging from .19 and below, but their difficulty levels are all highly acceptable which is within the range of .26 to .75. The marginal items includes one easy item (item 21 with difficulty level of .8); one difficult item (item 42 with difficulty level of .23); and twelve items of optimum difficulty (items 9, 15, 17, 19, 24, 35, 41, 43, 44, 52, 53, and 68) with difficulty indices ranging from .26. Table 1 shows the retained items in the final form of the Chemistry Achievement Test.

Table 1. Retained Items in the Chemistry Achievement Test

Former Item No.	Index of Difficulty	Index of Discrimination	New Item No.
2	0.5	0.32	1
3	0.7	0.53	2
4	0.59	0.44	3
6	0.48	0.72	4
7	0.39	0.46	5
8	0.69	0.5	6
9	0.52	0.21	7
10	0.5	0.44	8
12	0.53	0.32	9
14	0.62	0.63	10
15	0.53	0.25	11
16	0.67	0.53	12
17	0.43	0.25	13
18*	0.48	0.15	14
19	0.39	0.22	15
20	0.37	0.5	16
21	0.8	0.22	17
24	0.45	0.28	18
26	0.34	0.31	19
27	0.57	0.41	20

28	0.4	0.43	21
33	0.43	0.31	22
35	0.73	0.22	23
37	0.37	0.5	24
38	0.54	0.47	25
39	0.81	0.31	26
41	0.48	0.28	27
42	0.23	0.22	28
43	0.26	0.21	29
44	0.51	0.21	30
46	0.59	0.31	31
47	0.54	0.65	32
48	0.51	0.47	33
49*	0.37	0.19	34
50	0.4	0.37	35
51	0.47	0.62	36
52	0.4	0.25	37
53	0.58	0.22	38
54	0.5	0.44	39
57	0.21	0.31	40
58	0.39	0.4	41
59	0.29	0.35	42
60	0.48	0.65	43
63*	0.5	0.18	44
64	0.44	0.5	45

68	0.64	0.28	46
69	0.56	0.38	47
71	0.47	0.5	48
72*	0.4	0.13	49
73	.56	.38	50

* Items with poor discrimination index but with highly acceptable difficulty index which were considered for inclusion in the final form of the Chemistry Achievement Test.

Table 2 shows the discarded items in the Chemistry Achievement Test. These discarded items have difficulty indices below .20 (very difficult) and above .80 (very easy) and discrimination indices below .20.

Table 2. Discarded Items in the Chemistry Achievement Test

Item No.	Index of Difficulty	Index of Discrimination
1	0.89	0.16
5	0.17	-0.03
11	0.86	0.1
13	0.57	0.03
22	0.73	-0.09
23	0.18	0.37
25	0.7	0.22
29	0.19	0.06
30	0.14	0.22
31	0.35	-0.03
32	0.2	-0.03
34	0.37	0
36	.51	- .09

40	0.16	0
45	0.29	0.09
55	0.26	0.15
56	0.39	-0.16
61	0.29	-0.15
62	0.45	-0.03
65	0.44	-0.44
66	0.12	0
67	0.14	0.16
70	0.23	0.03
74	0.2	0.03
75	0.42	-0.03

Table 3 shows the content areas; skills measured; and the number, percentages and placement of items in the Chemistry Achievement Test.

Table 3. Item Content of the Chemistry Achievement Test

Cognitive,	CONTENT					Total	Item	Placement
Domain	Unit I	Unit II	Unit III	Unit IV	Unit V	%		
Knowledge		1,2	3	4	5	5	10	1-5
Comprehension	6	7	8		9	4	8	6-9
Application	10, 11	12, 13, 14,15	16, 17, 18	19, 20, 21	22	13	26	10-22
Analysis	23,24, 25,26	27,28, 29,30		31,32, 33,34,35	36,37, 38,39	17	34	23-39
Synthesis	40,41	42,43		44,45	46	7	14	40-46
Evaluation	47	48,49	50			4	8	47-50
No. of Items	10	15	6	11	8	50	100%	
Percent	20%	30%	12%	22%	16%	100%		

The instrument used in this study, the Chemistry Achievement Test, was composed of 50 items. These items were taken from five instructional units and were classified according to different levels of the cognitive domain, namely: 5 or 10 percent were knowledge level questions; 4 or 8 percent were comprehension level questions; 13 or 26 percent were application level questions; 17 or 34 percent were analysis level questions; 7 or 14 percent were synthesis level questions; and 4 or 8 percent were evaluation level questions. The cognitive levels ranged from simple to complex. These questions were taken from the different learning competencies in Chemistry for a whole school year.

Unit 1 deals with introductory concepts in Chemistry composed of three chapters. Unit 2 was about the concept of matter consisting of three chapters dealing with behavior of molecules, a view of the atom, and atoms in the periodic table. Unit 3 deals with why and how atoms combine. It was composed of two chapters dealing with bond formation and shape of molecules. Unit 4 deals with the factors affecting chemical reactions. It has three chapters, two of which deal with chemical activities and the other deals with chemical equilibrium. Unit 5 deals with how chemistry creates new technologies. It was composed of four chapters dealing with solutions; acids, bases, and salts; colloids; and life and carbon compounds.

Of the 5 knowledge level items, none was taken from Unit 1; items 1 and 2 were taken from Unit 2; item 3 was taken from Unit 3; item 4 was taken from Unit 4; and item 5 was taken from Unit 5. Of the 4 comprehension items, item 6 was taken from Unit 1; item 7 was taken from Unit 2; item 8 was taken from Unit 3; no item was taken from Unit 4; and item 9 was taken from Unit 5. Of the 13 application level items, items 10 and 11 were taken from Unit 1; items 12, 13, 14, and 15 were taken from Unit 2; items 16, 17, and 18 were taken from Unit 3; items 19, 20, and 21 were taken from Unit 4; and item 22 was taken from Unit 5. Of the 17 analysis level items, items 23, 24, 25, and 26 were taken from Unit 1; items 27, 28, 29, and 30 were taken from Unit 2; no item was taken from Unit 3; items 31, 32, 33, 34, and 35 were taken from Unit 4; and items 36, 37, 38, and 39 were taken from Unit 5. Of the 7 synthesis level items, items 40 and 41 were taken from Unit 1; items 42, and 43 were taken from Unit 2; no item qualifies from Unit 3; items 44 and 45 were taken from Unit 4; and item 46 was taken from Unit 5. Of the 4 evaluation items, item 47 was taken from Unit 1; items 48 and 49 were taken from Unit 2; and item 50 was taken from Unit 3. No item qualifies from Unit 4 and 5.

As per instructional unit, 10 or 20 percent of the items were taken from Unit 1; 15 or 30 percent of the items were taken from Unit 2; 6 or 12 percent of the items came from Unit 3; 11 or 22 percent of the items came from Unit 4; and 8 or 16 percent of the items were taken from Unit 5.

Unit 1 was composed of one comprehension item (item 6); two application items (items 10 and 11); four analysis items (items 23, 24, 25, and 26); two synthesis items (items 40 and 41); and one evaluation item (item 47). Unit 2 was composed of two knowledge level items (items 1 and 2); one comprehension item (item 7); four application items (items 12, 13, 14, and 15); four analysis items (items 27, 28, 29, and 30); two synthesis items (items 42 and 43); and two evaluation items (items 48 and 49). Unit 3 was composed of one knowledge level item (item 3); one comprehension item (item 8); three application items (items 16, 17, and 18); no analysis and synthesis items; and one evaluation item (item 50). Unit 4 was composed of one knowledge level item (item 4); no comprehension item; three application items (items 19, 20, and 21); five analysis items (items 31, 32, 33, 34, and 35); two synthesis items (items 44 and 45); and no evaluation item. Unit 5 was composed of one knowledge item (item 5); one comprehension item (item 9); one application item (item 22); four analysis items (items 36, 37, 38, and 39); one synthesis item (item 46); and none of evaluation item.

The data gathering procedures involved: (1) administration of the test to the public and private school examinees; and (2) checking and scoring the test. While, the data analysis procedure includes: (1) assigning and matching of test papers to the matched examinees by section and total score; (2) organizing data for every item into a three-way contingency table; (3) encoding data in the Statistical Analysis System (SAS) computer program; (4) DIF analyses for detecting and testing for differential item functioning between the matched examinees; and (5) eliminating the correct responses on the DIF items identified in the public/private matched group by each of the DIF methods, respectively; (6) retaining the unbiased items, herein referred to as the test version for each of the DIF methods, respectively; (7) assessment of the test versions' content validity, concurrent validity, and internal consistency reliability.

The four Contingency Table (CT) methods applied in the differential item functioning analysis 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 approach examines the likelihood or probability—of test takers from different groups with the same ability levels correctly responding to an item. An item is considered unbiased when all persons at a given ability level have an equal probability of correctly answering an item regardless of their group membership. The null 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 matched groups of examinees".

The Distractor Response Analysis examines the incorrect alternatives to a test item for differences in patterns of response among different subgroups of a population. The function of Distractor is to determine the significance of the differences between two or more group's response frequencies in the discrete categories of question distractors. The null hypothesis under test is that "there is no significant difference in proportions selecting distractors on the test items between the matched groups of examinees".

The Logistic Regression 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. In the LR analysis the predictor variables are the (a) score interval and school type for the public and private school examinees. The dependent variable is the odds of getting the item right. A significant score interval indicates that examinees with higher total score tend to score better in the examination. Likewise, a significant school type indicates that the odds of getting an item right are different between the matched examinees. The null 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".

The Mantel-Haenszel Statistic is a non-parametric contingency table procedure commonly used to perform statistical test for uniform DIF. The MH procedure is also used to estimate a ratio that yields a measure of effect size for evaluating the amount of DIF that is present. This ratio value is transformed to produce the Delta-MH (D-MH). A positive D-MH indicates DIF in favor of the focal groups, and a negative value signifies DIF in favor of the reference groups. The degrees of DIF in test items are labeled A, B, and C. The MH analysis 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 over all items after controlling for total test score between the matched groups of examinees".

The common measure of DIF is the statistical significance of the obtained chi square value for each item. A statistically significant chi square value indicates: a) difference in proportion attaining a correct response across total score categories for the Chi-Square approach; b) difference in proportions selecting distractors for the Distractor Response Analysis; c) the odds of getting the item right are different between two groups of interest for the Logistic Regression; and d) large Differential Item Functioning (DIF) effect for the Mantel-Haenszel Statistic.

To achieve statistical significance in all analyses, the computed chi-square value must be greater than the critical chi-square value of 3.84 and its associated probability should be less than the set alpha level of 0.05. The agreement of any two, three or all of the DIF approaches is indicated by their obtained measure of DIF. If any two, three or all of the four methods similarly obtained a statistically significant measure of DIF (chi square value) on an item or number of items, such methods were in agreement. If not, there is disagreement.

Table 4 shows the statistical criteria for eliminating DIF items.

Table 4. Statistical Criteria for Identifying DIF Items

DIF Methods	Focus of Analysis	Measure of DIF
Chi Square	Difference in proportion attaining a correct response across total score categories	Significance of chi square
Distractor Response Analysis	Difference in proportions selecting distractors	Significance of chi square
Logistic Regression	Odds of getting an item right	Significance of chi square

Mantel-Haenszel Statistics	Perform statistical test for evaluating the amount of DIF	Significance of chi square and large DIF effect

After eliminating the DIF items identified respectively by each of the DIF methods, the original test and the DIF-free versions were compared in terms of content validity, concurrent validity, and internal consistency reliability. The content validity of the original test and the DIF-free versions was estimated based on the remaining items after eliminating the correct responses on the DIF items identified by each of the DIF methods. The degree to which the items composed an adequate sample was based on a six-point scale. The six points are: Adequate (86-100%), Moderately Adequate (71-85%), Slightly Adequate (56-70%), Slightly Inadequate (41-55%), Moderately Inadequate (26-40%), and Inadequate (25% and below). The concurrent validity of the original and test versions were assessed by calculating the correlation coefficient between the examinees' scores and their grade point average in Science III. The Statistical Packages for the Social Sciences (SPSS) was used in calculating the Pearson correlation. The internal consistency reliability of the original test and its versions was determined by calculating the KR-20 Formula reliability coefficient.

3. Results

3.1 Differential Item Functioning Analysis

Table 5 shows the summary results from the DIF analysis to identify differential item functioning on the Chemistry Achievement Test between the public and the private school examinees.

Table 5. Summary Results of the DIF Analysis to Identify Differentially Functioning Items in the Public/Private Matched Examinees

Items	Concept/Skills Measured	X^2	DRA	LR	MH
		DIF Against			
1	gas property illustrated by garbage smell	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	Pvt	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 problem on Boyle's Law	Pub	Pub	Pub	Pub
19	solving for the molar mass of Fe2 O3	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
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		
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	
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
50	factor which causes the nails to rust	Pvt			

Note. Pub = Public; Pvt = Private.

X^2 Analysis. The chi-square analysis identified 13 items indicating DIF between the public and the private school examinees. Nine of which, items 1, 3, 5, 9, 19, 30, 31, 33, and 47 were potentially biased against the private school examinees. Whereas, four items, items 13, 16, 22, and 37 were potentially biased against the public school examinees.

DRA Analysis. The distractor response analysis 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 potentially biased against the private school examinees. Whereas, six, items 8, 13, 16, 22, 36, and 41 indicate DIF against the public school examinees.

LR Analysis. The LR analysis identified 22 items indicating 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 the 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 potentially biased against the public school examinees. In each of these items, the odds of getting the item right favored the private school examinees.

MH Analysis. The MH analysis showed that 22 of the 50 items displayed statistical bias. Of these 22 items, ten indicate DIF against the private 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 against the private school examinees. Whereas, the other twelve items namely, items 2, 8, 13, 14, 16, 22, 32, 36, 37, 40, 41, and 46 indicate DIF against the public school examinees. Each of these twelve items obtained a significant MH chi square value and a negative log odds ratio indicative of DIF against the public school examinees.

3.2 Content and Concurrent Validity, and Reliability Analyses of the Test Versions

After eliminating from the analyses the correct responses on the DIF items identified respectively by each of the four DIF techniques, each of the test versions was analyzed and compared in terms of content validity, concurrent validity, and internal consistency reliability. The test versions, which were labeled after each of the DIF methods used, refer to the DIF-free or retained items in the Chemistry Achievement test.

Table 6 shows the content validity of the test versions. They were: 100 percent for the original test; 74 percent or moderately adequate for the X^2; 68 percent or slightly adequate for the Distractor; and 56 percent or also slightly adequate for both LR and MH.

Table 6. Content Validity of Test Versions

Test Version	No. of Items Retained	Percent	Description
Original Test	50	100	Adequate
Chi Square	37	74	Moderately adequate
Distractor	32	64	Slightly adequate
Logistic Regression	28	56	Slightly adequate
Mantel-Haenszel	28	56	Slightly adequate

Table 7 shows the concurrent validity coefficients of the test versions. These are: 0.585 for the original test; 0.507 for the MH; 0.504 for the X^2; 0.499 for the LR; and 0.462 for the Distractor. All of the validity coefficients indicate significant positive and moderate relationship between the examinees' test scores and their GPA in Science III. Thus, the null hypothesis that there is no significant relationship between the examinees test scores and their grades in Science III is rejected in favor of the alternative hypothesis.

Table 7. Concurrent Validity of the Test Versions

Test Version	Validity Coefficient	Description
Original Test	0.585**	moderate relationship
Mantel-Haenszel	0.507**	moderate relationship
Chi Square	0.504**	moderate relationship
Logistic Regression	0.499**	moderate relationship

Distractor	0.462**	moderate relationship

** $p < 0.01$.

Table 8 shows the internal consistency reliability of the test versions. They are: 0.71 for the original test, 0.64 for the X^2, 0.62 for the Distractor; 0.57 for the LR, and also 0.57 for the MH.

Table 8. Internal Consistency Reliability of Test Versions

Test Version	No. of Items	Reliability Coefficient
Original Test	50	0.71
Chi Square	37	0.64
Distractor	32	0.62
Logistic Regression	28	0.57
Mantel-Haenszel	28	0.57

4. Discussion

4.1 Differential Item Functioning Analysis

The differential performance between the matched group of examinees can be attributed to: (1) discrepancies in the curriculum content of the public and private school investigated; (2) attraction of the matched examinees to the incorrect options, (3) disparity in exposure of the matched examinees to the information, concepts, vocabularies, or skills reflected in the content of the DIF items, (4) items which may reflect information and/or skills that was not experienced by the matched examinees; (5) ambiguities in the item stem, keyed response, or distractors, and (6) overly difficult reading level or inability of the matched examinees to comprehend or understand the concepts reflected on the DIF items.

The Logistic Regression and the Mantel-Haenszel Statistic yielded very similar results with respect to uniform Differential Item Functioning (DIF). The findings in the reference/focal group analysis deserve further comment. First, the number of items exhibiting DIF with both the LR and the MH procedures seems high. Apparently, both LR and MH are very sensitive than the other 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 were 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 produce using the MH Statistic.

The four methods for detecting DIF 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: (a) the procedure is not selecting item at random; and (b) 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. The LR and MH procedures appeared to have demonstrated the external validity evidence mentioned above.

4.2 Content Validity

The content validity of a test is the degree to which the items are a representative sample of the content area of the identified construct. Valuable validity evidence can be gained by examining the relationship between the content of the test and the construct or domain the test is designed to measure (Reynolds et al., 2006). Reynolds (1998) notes that validity evidence based on test content focuses on how well the test items sample the behaviors or subject matter the test is designed to measure. In a similar vein, Anastasi and Urbina (1997) note that validity evidence based on test content involves the examination of the content of the test to determine whether it provides a representative sample of the domain being measured. o assess content coverage, the overall test and the degree to which the items cover the specified domain must be rated (Reynolds et al., 2006).

The content coverage of the Chemistry Achievement Test and the degree to which the items cover the specified domain was examined. The test was composed of 50 items. These items were taken from five instructional units and were classified according to different levels of the cognitive domain, namely: 5 or 10 percent were knowledge level questions; 4 or 8 percent were comprehension level questions; 13 or 26 percent were application level questions; 17 or 34 percent were analysis level questions; 7 or 14 percent were synthesis level questions; and 4 or 8 percent were evaluation level questions. The cognitive levels ranged from simple to complex. These questions were taken from the different learning competencies in Chemistry for a whole school year.

X^2 Version. Of the five knowledge level items, two items (items 1 and 2) were retained, accounting for only 4 percent of the items in the original test, whereas, three items (items 1, 3, and 5) were discarded, accounting for 6 percent reduction in the number of original items. Of the four comprehension level items, three items (items 6, 7, and 8) were retained accounting for 6 percent of the number of original items, whereas, only one item (item 9) was discarded accounting for another 2 percent reduction in the original item pool. Of the thirteen application level items, 9 items (items 10 and 11 of Unit 1, items 12, 14, and 15 of Unit 2, items 17 and 18 of Unit 3, and items 20 and 21 of Unit 4) were retained, accounting for 18 percent of the original item pool, whereas, four items (item 13 of Unit 2, item 16 of Unit 3, item 19 of Unit 4, and item 22 of Unit 5) were discarded, accounting for another 8 percent reduction in the original item pool. Of the seventeen analysis level items, 13 items (items 23, 24, 25, and 26 of Unit 1; items 27, 28, and 29 of Unit 2; items 32, 34, and 35 of Unit 4; and items 36, 38, and 39 of Unit 5) were retained, accounting for 26 percent of the original items, whereas, four items (item 30 of Unit 2; items 31 and 33 of Unit 4; and item 37 of Unit 5) were discarded, accounting for another 8 percent reduction in the original item pool. Of the seven synthesis level items, all items (items 40 and 41 of Unit 1, items 42 and 43 of Unit 2; items 44 and 45 of Unit 4; and item 46 of Unit 5) were retained, accounting for 14 percent of the original item pool. Of the four evaluation level items, 3 items (items 48 and 49 of Unit 2 and item 50 of Unit 3) were retained, accounting for 6 percent retention of the original item pool, whereas, one item (item 47 of Unit 1) was removed, accounting for another 2 percent reduction in the original item pool.

With the exclusion of DIF items, the Chi Square version would be composed of 2 or 4 percent knowledge questions; 3 or 6 percent comprehension questions; 9 or 18 percent application questions; 13 or 26 percent analysis questions; 7 or 14 percent synthesis questions; and 3 or 6 percent evaluation questions. Thus, the Chi Square test version would only be composed of 37 or 74 percent of the items in the original test version. Based on the six-point scale in determining the degree to which the retained items composed an adequate sample, the chi square test version was moderately adequate (falling within the 71-85 percent range).

DRA Version. Of the five knowledge level items, two items (item 2 of Unit 2 and item 4 of Unit 4) were retained, accounting for 4 percent retention of the original items. Whereas, 3 items (item 1 of Unit 2; item 3 of Unit 3; and item 5 of Unit 5) were discarded, accounting for 6 percent reduction in the original number of items. Of the 4 comprehension items, two items (item 6 of Unit 1 and item 7 of Unit 2) were retained. These retained items accounted for 4 percent retention of the original test items. Whereas, two items (item 8 of Unit 3 and item 9 of Unit 5) were eliminated, accounting for another 4 percent reduction in the original test items. Of the 13 application items, 8 items (items 10 and 11 of Unit 1; items 12, 14, and 15 of Unit 2; items 17 and 18 of Unit 3, and item 20 of Unit 4) were retained, accounting for 16 percent retention of the original test items. Whereas, 5 items (item 13 of Unit 2; item 16 of Unit 3; items 19 and 21 of Unit 4; and item 22 of Unit 5) were discarded, accounting for another 10 percent reduction in the original item pool. Of the 17 analysis items, 13 items (items 23, 24, 25 and 26 of Unit 1; items 27, 28, and 29 of Unit 2; items 31, 32, and 34 of Unit 4; and items 37, 38, and 39 of Unit 5) were retained. These retained items accounted for 26 percent of the original item pool. Whereas, 4 items (item 30 of Unit 1; items 33 and 35 of Unit 4; and item 36 of Unit 5) were eliminated. These discarded items accounted for 8 percent reduction from the original pool of items. Of the 7 synthesis items, 5 items (item 40 of Unit 1, item 42 of Unit 2, items 44 and 45 of Unit 4, and item 46 of Unit 5) were retained. These retained items accounted for 10 percent of the original test items. Whereas, only 2 items (item 41 of Unit 1 and item 43 of Unit 2) were discarded, accounting for another 4 percent reduction from the original test items. Of the 4 evaluation items, 2 items (item 48 and 49 of Unit 2) were retained. These items accounted for 4 percent retention in the original test items. Whereas, 2 items (item 47 of Unit 1 and item 50 of Unit 3) were eliminated, accounting for another 4 percent reduction in the original pool of items.

After the exclusion of DIF items, the refined DRA test version would consist of 2 or 4 percent knowledge level questions; 2 or 4 percent comprehension questions; 8 or 16 percent application questions; 13 or 26 percent analysis questions; 5 or 10 percent synthesis questions; and 2 or 4 percent evaluation questions. Thus, the DRA version would only be composed of 32 retained items which accounts for 64 percent of the items in the original test version. Whereas, the 18 discarded items accounted for 36 percent reduction in the original item pool. Based on the six-point scale in determining the degree to which the retained items composed an adequate sample, the DRA version was slightly adequate (falling within the 56-70 percent range).

LR Version. Of the 5 knowledge level items, only one item (item 4 of Unit 4) was retained. It accounted for only 2 percent retention of the items in the original test. Whereas, 4 items were discarded, namely, items 1 and 2 of Unit 2; item 3 of Unit 3; and item 5 of Unit 5. These discarded items accounted for 8 percent reduction in the content of the LR test version. As to the 4 comprehension items, two items (items 6 and 7 of Units 1 and 2) were retained. They accounted for 4 percent retention of the original item pool. While two items (items 8 and 9 of Units 3 and 4) were eliminated, accounting for another 2 percent reduction in the number of items in the LR test version. Among the 13 application items, the 6 retained items were: item 11 of Unit 1, items 12 and 15 of Unit 2, item s 17 and 18 of Unit 3, and item 20 of Unit 4. These retained items accounts for 12 percent of the original test items. Whereas, the 7 discarded items were item 10 of Unit 1, items 13 and 14 of Unit 2, item 16 of Unit 3, and items 19 and 21 of Unit 4, and item 22 of Unit 5. These discarded items represent another 14 percent reduction in the content of this test version. Among the 17 analysis items, the 11 retained items were: items 23, 24, and 25 of Unit 1; items 27, 28, and 29 of Unit 2; items 31, 34, and 35 of Unit 4; and items 38 and 39 of Unit 5. These retained items accounted for 22 percent of the original test items. Whereas, the 6 excluded items were item 26 of Unit 1; item 30 of Unit 2; items 32 and 33 of Unit 4; and items 36 and 37 of Unit 5. These discarded items accounted for another 12 percent reduction in the content of the LR test version. Of the 7 synthesis items, 5 items were retained. These were items 42 and 43 of Unit 2; items 44 and 45 of Unit 4; and item 46 of Unit 5. These retained items accounted for 10 percent of the original test items. Whereas, 2 items were discarded, namely, items 40 and 41 of Unit 1. These eliminated items accounted for another 4 percent reduction in the content of the LR test version. Of the 4 evaluation items, 3 items were retained. These were items 48 and 49 of Unit 2, and item 50 of Unit 3. These represented 6 percent of the original test items. Whereas, one item, item 47, was discarded. It accounted for another 2 percent reduction in the content of the LR test version.

With the elimination of DIF items, the LR test version of the Chemistry Achievement Test would consists of: 1 or 2 percent knowledge questions; 2 or 4 percent comprehension questions; 6 or 12 percent application questions; 11 or 22 percent analysis questions; 5 or 10 percent synthesis questions; and 3 or 6 percent evaluation questions. Thus, the LR test version would only be composed of 28 items, a 56 percent retention of the items in the original test version. The discarded 22 items accounted for 44 percent reduction in the content of the LR test version. Based on the six-point scale in determining the degree to which the retained items composed an adequate sample, the LR version was slightly adequate (falling within the 56-70 percent range).

MH Version. Of the 5 knowledge level items, two items, namely, items 4 and 5 of Units 4 and 5, were retained. These retained items accounted for 4 percent of the items in the original test. Whereas, three items, items 1 and 2 of Unit 2, and item 3 of Unit 3 were discarded. These discarded items accounted for 6 percent reduction in the content of the MH test version. Of the four comprehension items, two items (item 6 of Unit 1 and item 7 of Unit 2) were retained. These items accounted for 4 percent of the items in the original test. Whereas, the other two items (items 8 of Unit 3 and item 9 of Unit 5) were excluded. These items accounted for another 4 percent reduction in the content of the MH test version. Of the 13 application items, 6 items (item 11 of Unit 1; items 12 and 15 of Unit 2; items 17 and 18 of Unit 3; and item 20 of Unit 4) were retained. The retained items accounted for 12 percent of the items from the original test. Whereas, 7 items (item 10 of Unit 1; items 13 and 14 of Unit 2; item 16 of Unit 3; item 19 and 21 of Unit 4, and item 22 of Unit 5) were excluded. These excluded items accounted for another 14 percent reduction from the original test items. Of the 17 analysis items, 11 items were retained. They were items 23, 24, and 25 of Unit 1; items 27, 28, and 29 of Unit 2; items 31, 34, and 35 of Unit 4; and items 38 and 39 of Unit 5. These retained items accounted for 22 percent of the original test items. Whereas, six items, item 26 of Unit 1; item 30 of Unit 2; items 32 and 33 of Unit 4; and items 36 and 37 of Unit 5 were eliminated. These discarded items represented 12 percent reduction in the content of the MH test version. Of the seven synthesis items, 4 were retained. These were items 42 and 43 of Unit 2, items 44 and 45 of Unit 4. These accounted for 8 percent of the items from the original test. Whereas, three items, items 40 and 41 of Unit 1 and item 46 of Unit 5 were discarded. These items accounted for 6 percent reduction in the content of the MH test version. Of the 4 evaluation items, 3 items, items 48 and 49 of Unit 2 and item 50 of Unit 3 were retained. These retained items accounted for 6 percent of the items from the original test. Whereas, item 47 was eliminated, accounting for 2 percent reduction in the number of items in the MH test version.

With the exclusion of DIF items, the MH version would consist of: 2 or 4 percent knowledge questions; 2 or 4 percent comprehension questions; 6 or 12 percent application questions; 11 or 22 percent analysis questions; 4 or 8 percent synthesis questions; and 3 or 6 percent evaluation questions. Thus, the MH test version would only compose of 28 items, a 56 percent retention of the original test items. The 22 or 44 percent of the original test items were excluded due to DIF. Based on the six-point scale in determining the degree to which the retained items composed an adequate sample, the MH version was slightly adequate (falling within the 56-70 percent range).

Thus, elimination of DIF items tend to diminished the content validity of a measuring instrument. That is, it lessens the number of item content and/or behavior of the domain being assessed in the test. The content validity of the test versions varies, ranging from slightly adequate to moderately adequate. The content validity appeared to be inversely proportional to the number of items eliminated. The higher the number of items excluded, the lower the content validity and vice versa. Thus, content validity is a function of the adequacy or inadequacy of the sample of items in a test.

4.3 Concurrent Validity

In this study, concurrent validity evidence was secured by examining the relationship between the examinees' test score and their Grade Point Average (GPA) in Science III, the criterion. Criterion is defined as a "measure of some attribute or outcome that is of primary interest. The criterion can be academic performance as reflected by the GPA or anything else that is of importance to the user of the test. Correlation coefficient is often used to examine the relationship between a predictor and a criterion, and in this context the correlation coefficient is referred to as a validity coefficient (Reynolds et al., 2006).

Similarly, though of varying magnitude, all test versions obtained positive and significant concurrent validity coefficients. The correlation between the examinees' test scores and their grade point average in Science III for each test version indicate positive and moderate relationship. The result of the concurrent validity analysis indicates that reduction in the examinees' test scores, due to exclusion of correct responses on the DIF items, tends to reduce the size of the concurrent validity coefficient. The concurrent validity is high if the examinees' test scores have high magnitude of correlation with their grade point average in Science III. On the other hand, the concurrent validity is low if the examinees' test scores have low magnitude of correlation with their grade point average in Science III.

Concurrent validity coefficient also indicates homogeneity of a group of test scores. The higher the correlation, the less homogeneous a group of scores, meaning the bigger the range of scores, the bigger the correlation coefficient will be. Whereas, the lower the correlation, the more homogeneous a group of scores, indicating that the smaller the range of scores, the smaller the correlation will be (Ferguson & Takane, 1989). The more homogeneous the test becomes the less valid it is.

4.4 Internal Consistency Reliability

The approach used in estimating the reliability of the original test and its versions was based on the formula developed by Kuder and Richardson (1937). This approach examines the consistency of responding to all the individual items and is derived from a single administration of the test. The most commonly used formula is known as the Kuder-Richardson formula 20 (KR-20). KR-20 is sensitive to measurement error due to content sampling and is also a measure of item heterogeneity. It is applicable when test items are scored dichotomously, that is, simply right or wrong, as 0 or 1 (Reynolds et al., 2006).

The KR-20 reliability analysis indicates that the higher the number of test items retained (DIF-free items), the higher the internal consistency reliability. Whereas, the lower the number of items retained, the lower was the internal consistency reliability. The reliability coefficient would be high if the items on the test have high inter-correlations with each other and are measure of much the same attribute. The reliability coefficient would be low if the items on the test have low inter-correlations. Items' inter-correlations are low, either because the items measure different attributes or because of the presence of error. A test may be made more reliable by increasing its length (Ferguson & Takane, 1989).

Thus, the study revealed that the higher the number of DIF items eliminated in a test, the lower was the content validity, concurrent validity, and the internal consistency reliability becomes, as it decreases the length or number of items of the test.

Possibly, the most obvious way to improve the reliability of measurement is simply to increase the number of items on a test. If we increase the number of items while maintaining the same quality as the original items, we will increase the reliability of the test (Reynolds et al., 2006).

In view of the aforementioned concept, the measurement qualities of a test can be maintained or enhanced by means of DIF correction. DIF correction could be done by replacing or revising DIF items and then re-administer the test and subject it anew to DIF analysis. In this technique, the required number of items in a test can be maintained and its reliability and validity could also be maintained and possibly enhanced.

5. Conclusions

The results of the differential item functioning analysis showed that there were statistically biased test items between the public and the private school examinees. A clear pattern shows that the biased items against the private school examinees were relatively easier items, mostly having difficulty indices of 0.5 and above. Whereas, biased items against the public school examinees were relatively difficult, mostly within the difficulty ranges of 0.5 and below.

There were agreement and disagreement among the DIF methods 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 methods 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 a high degree of correspondence between these two procedures.

Elimination of DIF items in a test tends to diminish its content validity, concurrent validity, and internal consistency reliability, as it decreases the length or number of items of the test.

6. Recommendations

The use of statistical methods in identifying test items indicating DIF is a relatively better kind of item analysis. This is so, because by subjecting test items to DIF 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 DIF items a valid, reliable, and fairer test would be made.

Educational evaluation practitioners should engage in DIF detection and may use Logistic Regression or Mantel-Haenszel Statistic for DIF correction, which means that identified DIF item or items must either be revised or replaced. Then, re-administer the test and subject it anew to DIF 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. DIF correction could maintain or improve the measurement qualities of a test such as its content validity, concurrent validity, and internal consistency reliability.

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

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

References

Camilli, G., & Shepard, L. (1994). *Methods for Identifying Biased Test Items* (Vol. 4). Sage Publications, Inc., California.

Ferguson, G. A., & Takane, Y. (1989). *Statistical Analysis in Psychology and Education* (6th ed.). New York: McGraw-Hill, Inc.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

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. https://dx.doi.org/10.1177/008124630703700104

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. https://dx.doi.org/10.1111/j.1745-3984.1995.tb00459.x

Osterlind, S. J. (1983). *Test Item Bias*. Sage Publications, Inc., California. https://doi.org/10.1111/j.1745-3984.1995.tb00459.x

Reynolds, C. R., Livingston, R. B., & Willson, V. (2006). *Measurement and Assessment in Education*. Pearson Education, Inc.

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. https://dx.doi.org/10.1177/014662169301700201

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement*, *27*, 361-370. https://dx.doi.org/10.1111/j.1745-3984.1990.tb00754.x