SAINSAB
ISSN 1511 5267 Vol. 17, 2014, pp 24-37

REGION RELATED DIFFERENTIAL ITEM FUNCTIONING ON NUMERICAL REASONING TEST

Aye Aye Myint & Cherry Zin Oo
Yangon Institute of Education
Myanmar
aamyint15259@gmail.com

Abstract

The main purpose of this study was to investigate the region related differential item functioning (DIF) by the use of numerical reasoning test which was adapted from the research of Aye Aye Myint (1997). To obtain the required data, a total of 1802 students (905 students from selected eleven high schools from Yangon city and 897 students from selected five high schools from Pathein city) participated in this study by using the stratified random sampling technique. Survey descriptive research method was used in this study. As a result of the BILOG-MG 3 programme, it was found that Items 2, 5, 10, 13 and 22 show uniform differential item functioning (DIF). So, the items may not be an equivalent measure of the same latent variable for both groups. Among the three categories of numerical reasoning ability, the inductive reasoning ability is the highest for Myanmar high school students. The application in this study provided good information not only for the differential item functioning (DIF) analysis but also guided the direction of numerical reasoning ability test.

Importance of the study

Psychological testing is a field characterized by the use of samples of behavior in order to assess psychological constructs (unobserved constructs), such as cognitive and emotional functioning, about a given individual. Psychological tests have a substantial impact on a variety of important decisions (Murphy & Davidsofer, 1988). The three domains that are most relevant to decision making are the domains of ability, interest and personality.

The technical term for the science behind psychological testing is psychometrics. There are many different kinds of psychometric test. The most common are numerical reasoning, verbal reasoning, and personality test. Numerical reasoning ability is the primary mental ability of Thurstone for intelligence (Benjamin et al. 1990, Murphy & Davidsofer, 1988). In this study, numerical reasoning ability tests assess the ability to use numbers in a logical and rational way. The questions measure the understanding of such things as number series, numerical transformations, the relationships between numbers and the ability to perform numerical calculations. The examiners need to work out how to get the answer, rather than just doing the necessary calculations. On the other hand, it is important that students who are equal on the measured construct also have an equal probability of answering the test items correctly. Group membership can also influence the chance of correctly answering items (differential item functioning). Test items are designed to provide information about the examinee. One potential threat to test score validity is item bias, which occurs when a test item unfairly favors one group of examinees over another, meaning that one group more often correctly answers the item. A biased item will exhibit differential item functioning (DIF).

Differential Item Functioning (DIF) occurs when examinees from different groups with equal knowledge exhibit different probabilities of success on an item. The differential item functioning (DIF) analysis is a procedure used to determine if test questions are fair and appropriate for assessing the knowledge of various groups. It is based on the assumption that test takers who have similar knowledge (based on total test scores) should perform in similar ways on individual test questions regardless of their sex, race, or region. This procedure was developed in 1986 by researchers from the Educational Testing Service (ETS) and is used with many ETS tests.

Objectives of the study

This study is conducted with the following objectives:

- To identify items that function differentially between Yangon city and Pathein city;
- To examine a possible interaction between region with respect to Differential Item Functioning (DIF);
- To determine which subgroup have the DIF items if any are in favor of ;
- To explore the differences of students' numerical reasoning ability by categories of numerical reasoning ability test by region, by types of school and by strata.

Review of related literature

The cognitive ability test is an assessment of a range of reasoning skills. Cognitive ability tests assess abilities involved in thinking (e.g., reasoning, perception, memory, verbal and mathematical ability, and problem solving). The numerical reasoning skill is the foundation of all other numerical abilities. This skill enables individuals to learn how to evaluate situations, how to select and apply strategies for problem-solving, how to draw logical conclusions using numerical data, how to describe and develop solutions, and to recognize when and how to apply the solutions (Rust, 2006).

Item Response Theory (IRT) is a general statistical theory about examinee item and test performance and how performance relates to the abilities that are measured by the items in the test (Hambleton & Jones, 1993). Item response theory (IRT) is a collection of models that provide information about the properties of items and the scales they comprise through the analysis of individual item responses (Crocker & Algina, 1986). An item response theory (IRT) model is ideally suited for the detection of differential item functioning (DIF) in examining the validity of a test or questionnaire.

Dorans and Holland (1993) defined differential item functioning as "differences in item functioning *after* groups have been matched with respect to the ability or attribute that the item purportedly measures" (p. 37, italics in original). The concept of differential item functioning (DIF) refers to different performance of an item for members of different groups that are equal in the ability which are measured by the test (e.g. numerical reasoning ability). Differential item functioning (DIF) refers to a difference in item performance between two comparable groups of examinees, that is, groups that are matched with respect to the construct being measured by the test (Pour & Ghafar, 2009). Kristjansson *et al*. (2005) also distinguish between uniform DIF and non-uniform DIF. For dichotomous (binary) items, uniform DIF occurs when the item is more difficult at all ability levels for one group than for the other. Wa Kivilu (2010), by way of contrast, states that there is uniform DIF when there is dependence on group membership but no interaction between score category and group membership (indicated by two parallel ICCs). According to Kristjansson *et al*. (2005), non-uniform DIF occurs when there is an interaction between ability level and group so that the item is more difficult, for example, for one group at lower levels of ability but more difficult for the other groups at higher levels of ability.

Sample

The townships in Yangon city and Pathein city were stratified on the basis of geographical, economical, and social characteristics. These townships are classified into three strata, that is, Inner City, Inner Suburb, and Outer Suburb. Secondly, a sample of schools from each stratum was selected. The schools were divided into three groups as high group, the middle group and the low group according to the percentages of successful candidates of matriculation examination results. The number of students participated in test administration by school from Yangon city and Pathein city are shown in table 1.

Table 1 Distribution of sample by stratum, types of school and region

	Types of School			
	High	Middle	Low	Total
Yangon	288 (31.82%) [4]*	467 (51.6%) [4]*	150 (16.75%) [3]*	905 [11]*
Pathein	227 (25.3%) [1]*	453 (50.51%) [2]*	217 (24.19%) [1]*	897 [4]*
	Stratum			
	Inner city	Inner suburb	Outer suburb	Total
Yangon	131 (14.48%) [2]*	188 (20.77%) [3]*	586 (64.75%) [6]*	905 [11]*
Pathein	227 (25.3%) [1]*	453 (50.51%) [2]*	217 (24.19%) [1]*	897 [4]*

*Number of schools

Method

The numerical reasoning test developed by Aye Aye Myint (2000) was adapted and applied to measure the students' level of numerical reasoning ability. It included three kinds of item format. Type A contained 10 word problems, which involved simple calculations demanding one or more than one step or operations, which emphasis problem and algorithmic thinking.

Type B consisted of 4 numerical equations which involved the determination of the value of unknown digits. Each item was an equation in which one or more digits have been replaced by symbols. The symbol □ has to be a digit between 1 and 9 inclusive. Several logical and numerical steps must be taken in order to solve each equation.

Type C consisted of 11 items of finite number and series and numeric inference. One has to bring about a certain computational rule between the consecutive numbers along the series or in the three or four pairs of numbers by using inductive reasoning ability.

All of test items of numerical reasoning test (NRT) were multiple choice items. The test items were written in Myanmar Language, the mother tongue of the students. Students completed the 25 items multiple choice test in 45 minutes under actual test administration. Dichotomously scoring method was used for the free response items. Test data analysis was made using SPSS 16.0, BILOG-MG 3, and Microsoft Excel.

Findings and Discussion

Item parameters and ability parameters were estimated with the Differential Item Functioning (DIF) by using BILOG-MG 3 Software package. As an output, the program provides Phase I output, Phase II output and Phase III output. Phase I output includes test and item identification and classical item statistics separately for each group. Each subtest is calibrated separately in Phase II output. Phase III output is created, but no scoring is performed in the case of a DIF analysis.

Results of item statistics for groups

For completeness, Both the Pearson and biserial item-test correlations are shown. The reason for reporting these correlations separately for each group is that the appearance of large discrepancies between groups for a given item would suggest that the assumption of a common slope is untenable.

Table 2 Item statistics for Yangon city

ITEM	NAME	#TRIED	#RIGHT	PEARSON	BISERIAL
1	ITEM0001	905.0	781.0	0.282	0.442
2	ITEM0002	905.0	776.0	0.276	0.428
3	ITEM0003	905.0	671.0	0.282	0.381
4	ITEM0004	905.0	592.0	0.443	0.571
5	ITEM0005	905.0	515.0	0.400	0.504
6	ITEM0006	905.0	620.0	0.443	0.579
7	ITEM0007	905.0	355.0	0.462	0.587
8	ITEM0008	905.0	333.0	0.411	0.526
9	ITEM0009	905.0	265.0	0.337	0.445
10	ITEM0010	905.0	109.0	0.187	0.304
11	ITEM0011	905.0	756.0	0.313	0.469
12	ITEM0012	905.0	558.0	0.408	0.519
13	ITEM0013	905.0	283.0	0.481	0.630
14	ITEM0014	905.0	432.0	0.392	0.491
15	ITEM0015	905.0	649.0	0.392	0.522
16	ITEM0016	905.0	703.0	0.424	0.591
17	ITEM0017	905.0	470.0	0.498	0.624
18	ITEM0018	905.0	314.0	0.497	0.641
19	ITEM0019	905.0	373.0	0.457	0.578
20	ITEM0020	905.0	633.0	0.399	0.525
21	ITEM0021	905.0	521.0	0.381	0.481
22	ITEM0022	905.0	252.0	0.384	0.513
23	ITEM0023	905.0	299.0	0.556	0.721
24	ITEM0024	905.0	394.0	0.423	0.533
25	ITEM0025	905.0	446.0	0.433	0.543

ISSN 1511 5267 Vol. 17, 2014, pp 24-37

Table 3 Item statistics for Pathein city

ITEM	NAME	#TRIED	#RIGHT	PEARSON	BISERIAL
1	ITEM0001	897.0	751.0	0.361	0.541
2	ITEM0002	897.0	808.0	0.288	0.493
3	ITEM0003	897.0	679.0	0.339	0.464
4	ITEM0004	897.0	568.0	0.518	0.663
5	ITEM0005	897.0	646.0	0.227	0.303
6	ITEM0006	897.0	622.0	0.292	0.383
7	ITEM0007	897.0	383.0	0.483	0.609
8	ITEM0008	897.0	296.0	0.289	0.375
9	ITEM0009	897.0	260.0	0.174	0.230
10	ITEM0010	897.0	0.0	0.000	0.000
11	ITEM0011	897.0	556.0	0.436	0.555
12	ITEM0012	897.0	493.0	0.444	0.558
13	ITEM0013	897.0	165.0	0.287	0.418
14	ITEM0014	897.0	367.0	0.512	0.648
15	ITEM0015	897.0	741.0	0.525	0.775
16	ITEM0016	897.0	697.0	0.585	0.816
17	ITEM0017	897.0	460.0	0.591	0.740
18	ITEM0018	897.0	425.0	0.516	0.647
19	ITEM0019	897.0	525.0	0.455	0.576
20	ITEM0020	897.0	708.0	0.471	0.666
21	ITEM0021	897.0	654.0	0.441	0.592
22	ITEM0022	897.0	535.0	0.366	0.463
23	ITEM0023	897.0	423.0	0.637	0.799
24	ITEM0024	897.0	469.0	0.503	0.631
25	ITEM0025	897.0	591.0	0.551	0.711

Table 4 Item statistics for multiple regions

ITEM	NAME	#TRIED	#RIGHT	PEARSON	BISERIAL
1	ITEM0001	1802.0	1532.0	0.318	0.487
2	ITEM0002	1802.0	1584.0	0.284	0.461
3	ITEM0003	1802.0	1350.0	0.310	0.422
4	ITEM0004	1802.0	1160.0	0.477	0.612
5	ITEM0005	1802.0	1161.0	0.323	0.415
6	ITEM0006	1802.0	1242.0	0.367	0.480
7	ITEM0007	1802.0	738.0	0.474	0.599
8	ITEM0008	1802.0	629.0	0.345	0.444
9	ITEM0009	1802.0	525.0	0.253	0.336
10	ITEM0010	1802.0	109.0	0.102	0.202
11	ITEM0011	1802.0	1312.0	0.339	0.454
12	ITEM0012	1802.0	1051.0	0.417	0.527
13	ITEM0013	1802.0	448.0	0.368	0.503
14	ITEM0014	1802.0	799.0	0.441	0.555
15	ITEM0015	1802.0	1390.0	0.456	0.633
16	ITEM0016	1802.0	1400.0	0.502	0.700

17	ITEM0017	1802.0	930.0	0.541	0.679
18	ITEM0018	1802.0	739.0	0.511	0.646
19	ITEM0019	1802.0	898.0	0.461	0.577
20	ITEM0020	1802.0	1341.0	0.437	0.592
21	ITEM0021	1802.0	1175.0	0.414	0.534
22	ITEM0022	1802.0	787.0	0.373	0.469
23	ITEM0023	1802.0	722.0	0.600	0.761
24	ITEM0024	1802.0	863.0	0.467	0.586
25	ITEM0025	1802.0	1037.0	0.494	0.623

For the students in Yangon city, the result can be seen that the Pearson correlation coefficient is 0.187 to 0.557 (see table 2). For the students in Pathein city, the Pearson correlation coefficient is 0.000 to 0.591 (see table 3). For both regions, the correlation coefficient is 0.102 to 0.541 (see table 4). The result of the direction of the correlation coefficient is positive. It can be pointed that the correlation coefficient is high among items. The value of the biserial correlation coefficient between items for Yangon city is 0.304 to 0.721. For Pathein city, the value of the biserial correlation coefficient between items is 0.000 to 0.799. For both regions, that value is 0.202 to 0.761. The result can be seen that the correlation between numerical reasoning ability and total test score is high.

Investigation of the model for group differential item functioning

Parameter estimates for the variants are computed with respect to the latent dimension determined by main items.

Table 5 Item parameters for Yangon city

Item	Intercept S.E	Slope S.E	Threshold S.E	Asymptote S.E
ITEM0001	1.390 0.066*	0.799 0.079*	-1.740 0.139*	0.000 0.000*
ITEM0002	1.333 0.084*	0.751 0.069*	-1.775 0.143*	0.000 0.000*
ITEM0003	0.802 0.052*	0.646 0.060*	-1.241 0.108*	0.000 0.000*
ITEM0004	0.690 0.053*	1.045 0.071*	-0.660 0.049*	0.000 0.000*
ITEM0005	0.250 0.045*	0.534 0.050*	-0.469 0.083*	0.000 0.000*
ITEM0006	0.615 0.051*	0.635 0.056*	-0.969 0.088*	0.000 0.000*
ITEM0007	-0.199 0.046*	0.806 0.053*	0.247 0.061*	0.000 0.000*
ITEM0008	-0.307 0.045*	0.546 0.043*	0.562 0.096*	0.000 0.000*
ITEM0009	-0. 541	0.414	1.307	0.000

	0.047*	0.039*	0.164*	0.000*
ITEM0010	-1.240 0.065*	0.332 0.046*	3.739 0.497*	0.000 0.000*
ITEM0011	1.255 0.068*	2.365 0.073*	-0.531 0.022*	0.000 0.000*
ITEM0012	0.483 0.050*	0.827 0.063*	-0.584 0.057*	0.000 0.000*
ITEM0013	-0.476 0.051*	0.780 0.063*	0.609 0.085*	0.000 0.000*
ITEM0014	0.084 0.045*	0.885 0.062*	-0.095 0.050*	0.000 0.000*
ITEM0015	1.121 0.061*	1.463 0.093*	-0.766 0.042*	0.000 0.000*
ITEM0016	1.559 0.082*	1.759 0.116*	-0.886 0.040*	0.000 0.000*
ITEM0017	0.340 0.052*	1.314 0.084*	-0.259 0.036*	0.000 0.000*
ITEM0018	-0.329 0.051*	1.011 0.073*	0.326 0.059*	0.000 0.000*
ITEM0019	-0.112 0.048*	0.932 0.064*	0.121 0.054*	0.000 0.000*
ITEM0020	0.907 0.058*	1.175 0.078*	-0.772 0.048*	0.000 0.000*
ITEM0021	0.395 0.048*	0.926 0.062*	-0.427 0.049*	0.000 0.000*
ITEM0022	-0.606 0.050*	0.672 0.049*	0.901 0.101*	0.000 0.000*
ITEM0023	-0.352 0.057*	1.589 0.096*	0.222 0.040*	0.000 0.000*
ITEM0024	-0.032 0.047*	0.953 0.058*	0.034 0.050*	0.000 0.000*
ITEM0025	0.197 0.049*	1.127 0.072*	-0.175 0.041*	0.000 0.000*

* Standard Error

Table 6 Item parameters for Pathein city

Item	Intercept S.E	Slope S.E	Threshold S.E	Asymptote S.E
ITEM0001	1.291	0.799	-1.615	0.000

	0.073*	0.079*	0.124*	0.000*
ITEM0002	1.634 0.084*	0.751 0.069*	-2.176 0.162*	0.000 0.000*
ITEM0003	0.863 0.055*	0.646 0.060*	-1.336 0.114*	0.000 0.000*
ITEM0004	0.592 0.057*	1.045 0.071*	-0.566 0.048*	0.000 0.000*
ITEM0005	0.689 0.047*	0.534 0.050*	-1.292 0.128*	0.000 0.000*
ITEM0006	0.641 0.048*	0.635 0.056*	-1.009 0.094*	0.000 0.000*
ITEM0007	-0.130 0.047*	0.806 0.053*	0.161 0.059*	0.000 0.000*
ITEM0008	0.416 0.045*	0.546 0.043*	0.763 0.100*	0.000 0.000*
ITEM0009	-0.528 0.045*	0.414 0.039*	1.275 0.159*	0.000 0.000*
ITEM0011	0.480 0.054*	2.365 0.073*	-0.203 0.858*	0.000 0.000*
ITEM0012	0.250 0.047*	0.827 0.063*	-0.303 0.021*	0.000 0.000*
ITEM0013	-1.010 0.059*	0.780 0.063*	1.295 0.056*	0.000 0.000*
ITEM0014	-0.187 0.049*	0.885 0.062*	0.211 0.114*	0.000 0.000*
ITEM0015	1.733 0.103*	1.463 0.093*	-1.185 0.057*	0.000 0.000*
ITEM0016	1.678 0.102*	1.759 0.116*	-0.954 0.051*	0.000 0.000*
ITEM0017	0.204 0.053*	1.314 0.084*	-0.155 0.043*	0.000 0.000*
ITEM0018	0.031 0.048*	1.011 0.073*	-0.031 0.041*	0.000 0.000*
ITEM0019	0.390 0.050*	0.932 0.064*	-0.418 0.047*	0.000 0.000*
ITEM0020	1.311 0.074*	1.175 0.078*	-1.116 0.051*	0.000 0.000*
ITEM0021	0.898 0.059*	0.926 0.062*	-0.969 0.060*	0.000 0.000*

ITEM0022	0.356 0.045*	0.672 0.049*	-0.530 0.065*	0.000 0.000*
ITEM0023	0.053 0.058*	1.589 0.096*	-0.033 0.070*	0.000 0.000*
ITEM0024	0.188 0.048*	0.953 0.058*	-0.197 0.036*	0.000 0.000*
ITEM0025	0.721 0.059*	1.127 0.072*	-0.640 0.051*	0.000 0.000*

* Standard Error

It was found that item discrimination parameters (slope) range from 0.414 to 2.365 for both regions and the mean of a values is 0.971. Therefore, it is concluded by a consideration of their discrimination powers, the items are good items to provide appropriate discrimination of the test. From table 5, the variability of item difficulty parameters (threshold) for group 1 (Yangon city) ranges from -1.775 to 3.739 and the mean of b value is -0.131. From table 6, the variability of item difficulty parameters (threshold) for group 2 (Pathein city) ranges from -2.176 to 1.295 and the mean of b value is -0.439. It was found that 60% for group 1 and 76% for group 2 of the items are with negative value of b values (item difficulty parameter), and thus, it is concluded that the test is relatively easy. There is no asymptote value because in the DIF analysis, there is no scoring output. This test is based on dichotomously scoring items.

Table 7 Threshold Means

Group	Adjustment
Yangon city	0.000
Pathein city	-0.167

The mean threshold of the Pathein city is 0.167 below that of the Yangon city as shown in table 7. It may be said that the items were more easily answered by students in Pathein city.

Table 8 Estimation of the threshold differences for both regions

Item	Group 2 – 1	Item	Group 2 - 1	Item	Group 2 – 1
ITEM0001	0.291 0.186*	ITEM0010	-3.739 0.497*	ITEM0019	-0.372 0.072*
ITEM0002	-0.234 0.216*	ITEM0011	0.495 0.858*	ITEM0020	-0.177 0.070*
ITEM0003	0.071 0.157*	ITEM0012	0.448 0.060*	ITEM0021	-0.376 0.078*
ITEM0004	0.261 0.069*	ITEM0013	0.852 0.102*	ITEM0022	-1.264 0.120*
ITEM0005	-0.656 0.152*	ITEM0014	0.472 0.125*	ITEM0023	-0.088 0.081*
ITEM0006	0.126 0.129*	ITEM0015	-0.252 0.071*	ITEM0024	-0.064 0.062*
ITEM0007	0.081 0.085*	ITEM0016	0.099 0.065*	ITEM0025	-0.298 0.066*

ITEM0008	0.368 0.139*	ITEM0017	0.270 0.056*
ITEM0009	0.136 0.228*	ITEM0018	-0.190 0.072*

* Standard Error

By observing the table 8, it was most certain that the threshold differences between Yangon city and Pathein city. Yangon city have 13 items that have more threshold values than Pathein city. Pathein city have 12 items that have more threshold value than Yangon city. This result pointed out that the students in Yangon city can't answer easily in Items 2, 5, 10, 15, 18, 19, 20, 21, 22, 23, 24 and 25. The students in Pathein city can't answer easily in other Items.

Results of Information Curves for Both Cities

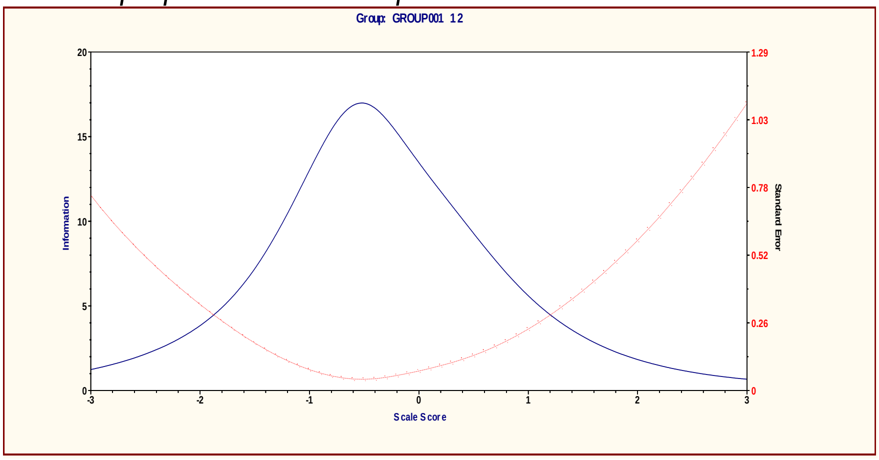

Figure 1 Information curve for Yangon city

Figure 1illustrates that the test has smaller standard errors across the ability scale from -1.9 to +1.2 and larger standard error at the low and high ends of the scale. The maximum amount of information I (θ) = 15.3 is at θ = -0.5. Ability estimates are more precise across the ability scale from -1.9 to +1.2 than at the low and high ends of the scale. Therefore, it was concluded that this test could be suitable for examinees whose numerical reasoning ability was θ = -0.5.

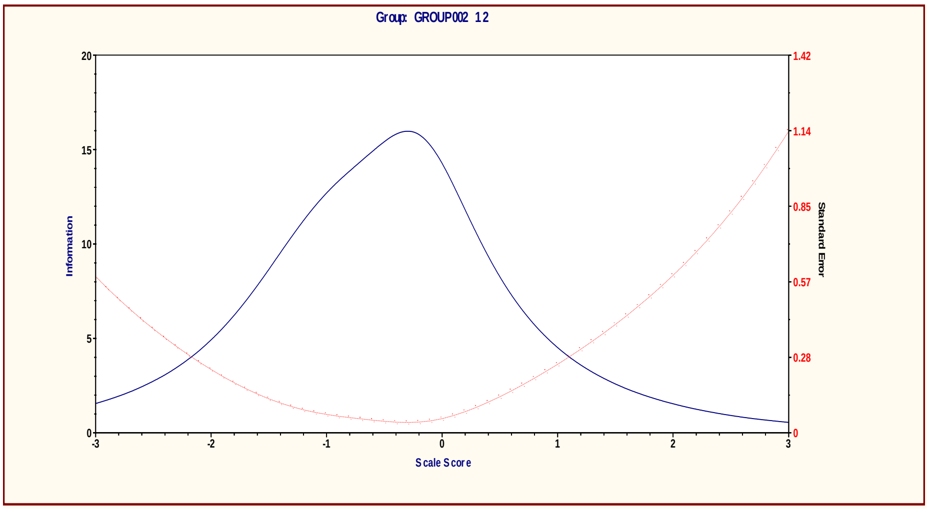

Figure 2 Information curve for Pathein city

According to Figure 2, the test has smaller standard errors across the ability scale from -2.1 to +1.1 and larger standard error at the low and high ends of the scale. The maximum amount of information I (θ) = 15.6 is at θ = -0.3. Ability estimates are more precise across the ability scale from -2.1 to +1.1 than at the low and high ends of the scale. Therefore, it was concluded that this test could be suitable for examinees whose numerical reasoning ability was θ = -0.3.

Items 2, 5, 10, 13 and 22 demonstrate differential item functioning (DIF). This type of DIF is known as uniform DIF because their item characteristic curves do not across. So the items may not be and equivalent measure of the same latent variable for both groups. It may be conclude that items 2, 5and 22 favor Pathein city and items 10 and 13 favor Yangon city.

Differences of numerical reasoning ability in three categories of numerical reasoning ability

Table 9 Descriptive statistics for each category of numerical reasoning ability

Each category of numerical reasoning ability	**Mean**	**Mean %**	**Std. Deviation**
Problem & Algorithmic thinking	5.56	55.6%	2.151
Logical reasoning	2.00	50%	1.285
Inductive reasoning	6.25	56.82%	3.209
Total	13.81		5.405

Based on the descriptive statistics shown in table 9, the mean % of the students' inductive reasoning is the highest of the three categories of numerical reasoning ability. It may be said that the high school students have the high inductive reasoning ability. So, the students can draw inferences from observations. It was observed that logical reasoning of high school students is weak among numerical reasoning abilities.

Comparison of Numerical Reasoning Ability Test by Region

Table 10 Independent sample t- test by region for each category of numerical reasoning ability test

Variable	t	df	Sig (2-tailed)	Mean difference
Problem & Algorithmic thinking	-.268	1800	.789	-.027
Logical reasoning	8.133	1800	.000	.484
Inductive reasoning	-9.019	1800	.000	-1.334

It was found that there were significant differences between Yangon city and Pathein city in logical reasoning and inductive reasoning. Therefore, this result pointed out the logical reasoning of students from Yangon city was better than that of students from Pathein city. In inductive reasoning, the students in Pathein city were better than the students in Yangon city.

School Wise Analysis and Results

Table 11 Multiple comparisons for numerical reasoning ability by types of school

Variable	(I) type of school	(J) type of school	Mean Difference (I-J)	Std. Error	Sig.
Problem & Algorithmic thinking	High	Middle	1.226*	.115	.000
	High	Low	1.520*	.141	.000
Logical reasoning	High	Middle	.924*	.067	.000
	High	Low	1.080*	.082	.000
Inductive reasoning	High	Middle	1.589*	.174	.000
	High	Low	1.230*	.213	.000

Note. * The mean difference is at 0.01 level.

In each category of numerical reasoning ability, the students from high school level have higher numerical reasoning ability than students from middle and low school levels. It was found that there was no significant different between middle and low school levels.

Stratum Wise Analysis and Results

Table 12 Multiple comparisons for numerical reasoning ability by strata

Variable	(I) Strata	(J) Strata	Mean Difference (I-J)	Std. Error	Sig.
Problem & Algorithmic thinking	Inner city	Inner suburb	1.200*	.139	.000
	Inner city	Outer suburb	1.372*	.134	.000
Logical reasoning	Inner city	Inner suburb	1.099*	.082	.000
	Inner suburb	Outer suburb	-.370*	.065	.000
	Inner city	Outer suburb	.730*	.079	.000
Inductive reasoning	Inner city	Inner suburb	1.901*	.209	.000
	Inner city	Outer suburb	1.517*	.201	.000

Note. * The mean difference is at 0.01 level.

It means that the students from inner city have higher numerical reasoning ability than students from inner suburb and outer suburb in each category. In logical reasoning ability, the students from outer suburb have high logical reasoning ability than that of the students from inner suburb. There was no significant difference between students' problem and algorithmic thinking and inductive reasoning among inner suburb and outer suburb.

Conclusion

In the examination of stability of the test, the test-retest reliability coefficient was found to be 0.81. An examination of the internal consistency reliability of the numerical reasoning test by Cronbach's alpha indicated that there was a high coefficient alpha value of 0.85 for both regions. Items 2, 5, 10, 13, and 22 show uniform DIF. So, the items may not be an equivalent measure of the same latent variable for both groups. Items 2, 5, 22 were items that display differential item functioning (DIF) favoring Pathein city for the comparison between Yangon city and Pathein city. Among 3 items, item 2 and 5 were items that involve simple calculations demanding one or more than one step or operations. Algorithmic thinking that involve mathematical abilities and the ability to solve routine problems has to be used to solve problem. The percentage of Pathein city that answered these two items correctly (90.17% and 58.71%) were higher than that of (85.75% and 56.91%). Item 22 was an item of finite number and series and numeric inference. Inductive processes and numerical steps must both be used to find out the answer. The percentage of Pathein city that answered these items correctly (60.79%) was higher than that of Yangon city (27.85%). The other item 10 and item 13 were items that display DIF favoring Yangon city for all the overall comparison between Yangon city and Pathein city. Item 10 was an item like item 2 and item 5. The correct answer was obtained by 12.84% Yangon city. There were no students who got the right answer to item 10 by Pathein city. Another item that displays DIF from Yangon city was item 13. Item 13 was an item of numerical equation which involve the determination of the value of unknown digits. The percentage of students from Yangon city that answered this item correctly (31.27%) was higher than that of students from Pathein city (20.1%). So, the results of this study indicated the presence of five items showing region DIF.

According to Lord (1980), the item information functions and the test information function play an important role and give a sound basis in test construction for selecting the items to fulfill the test requirements. Therefore, the item information functions of the 25 items and the test information function of the numerical reasoning test were obtained and investigated for both regions. Hence, the test provides the desired precise ability estimates across the ability scale.

Based on the descriptive statistics, the inductive reasoning ability of high school students is higher than the problem and algorithmic thinking and logical reasoning ability of high school students. Finally, this study investigated the differences exist among the strata and schools, and between region in the performance on the numerical reasoning test. Therefore, it is very important to detect, revise, or delete DIF items from the highly selective tests. The conclusions of the numerical reasoning test are particularly appropriate to inform differential item functioning (DIF) analysis in general. When one student in one country made a further study in other country, only differential item functioning (DIF) method can be used to show the ability and achievement of the students by comparing other students in other countries. The investigating of the comparison should be conducted not only region but also gender using DIF. The research should be performed not only at the global level but also within relevant subgroups.

So, the development of tests using differential item functioning (DIF) can be considered as a way to uplift the standard of Education as they are believed to help find whether solutions for more efficient teaching and learning.

References

Anastasi, A. (1997). *Psychological testing* (7th ed.). New Jersey: Pretiece-Hall, Inc.

Angoff, W. H. (1993). Perspective on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp.3-23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Aye Aye Myint. (1997). Investigation of the numerical reasoning ability of Myanmar high school students. *Bulletin of the Graduate School of Education, The University of Tokyo*, 37, 163-176.

Aye Aye Myint. (2000). *Development of numerical reasoning ability test and the vocabulary comprehension ability test for Myanmar high school students.* PhD thesis, Tokyo.

Benjamin, Jr. L., & Hopkins, J. R., & Nation, J. R. (1990). *Psychology*. (2nd ed). Macmillan Publishing Company, Macmillan, Inc.

Crocker, L. J., & Aligna, J. (1986). *Introduction to classical and modern test theory*. FL: Holt, Rinehart and Winston.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel –Haenzel and standardization. In P. W. Holland, & H. Wainer (Eds.), *Differential Item Functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gazzaniga, M. S. & Heatherton, T. F. (2006). *Psychological science* (2nd ed.). New York: W. W. Nortan & Company.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Items Instructional Topics in Educational Measurement.*

Kristjansson, E., Aylesworth, R., Mcdowell, I., & Zumbo, B.D. (2005). A comparison of four methods for detecting Differential Item Functioning in ordered response items. *Educational and Psychological Measurement*, *65*(6), 935–953. http://dx.doi.org/10.1177/0013164405275668

Lord, M. F. (1980). *Applications of item response theory to practical testing problems.* USA: Lawrence Erlbaum Associates, Broadway, Hillsdile, NJ.

Murphy, K. R., & Davidshofer, C. O. (1998). *Psychological Testing* (4th ed.). New Jersy: Pretiece-Hall, Inc.

Pour, I. M. & Ghafar, M. A. (2009). *The analysis of Iran universities 2003-2004 entrance examination to detect biased items*. Journal Tednologi, 50, pp.21-27.

Rust, J. (2006) *Advanced Numerical Reasoning Appraisal (ANRA) Manual.* NCS Pearson, Inc.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs, 1*.

Wa Kivilu, J.M. (2010). Determination of Differential Bundle Functioning (DBF) of numeracy and literacy tests administered to Grade 3 learners in South Africa. *South African Journal of Psychology*, *40*(3), 308–317.