A Comparison of the Classical Test Theory and the Multidimensional Item Response Theory total scores in a high scoring group of students on a Mathematics Multiple Choice Examination

Patt D. Sealy and *Smail Mahdi

Department of Mathematics Computer Science and Physics, Faculty of Pure and Applied Science, Cave Hill Campus, Barbados

Classical Test Theory was used to score a sample of 3000 examinees on a Mathematic Multiple Choice Examinations. Four scores were assigned to each examinee: a total correct score, which is the sum of all the items answered correctly and a score on each of three profile dimensions assigned by test constructors. Students who got 57% or more were selected for rescoring using MIRT; Bayes' estimation procedure implemented in the program TESTFACT was used. Since MIRT scores are Z-scores, each MIRT score was converted to a T-score for easy interpretation. Using MIRT, the 30 top students received "A's" and "A+'s"; whereas, under CTT, each examinee in this group got a straight grade A. When the total correct score was examined in the 30 students from the bottom of this high scoring group, all obtained a passing grade using a CTT scoring technique but a failing grade using a MIRT procedure. If a test is determined to be multidimensional a multidimensional scoring procedure should be used to estimate the student's ability; in addition, a passing or failing grade should be based on the scores awarded on each dimension and not solely on a Total Score; because, a total score ignores the individual's areas of strength in the subject and this could lead to incorrect interpretation of the examinee's ability.

Keywords: Dimensionality, Classical Test Theory Total Score, Multidimensionality, Latent trait, TESTFACT, Item Factor Analysis, MIRT score, T-score.

2000 AMS Subject Classification: 97D60; 97Cxx; 97K80.

Introduction

Walker and Beretvas (2000) argue that the primary reason for examining students in mathematics is to quantify the level of their mathematical ability. They further argue that since no measurement is free of measurement error, any one-dimensional procedure used to measure a multidimensional test may exacerbate this measurement error.

In solving Mathematics problems, students are required to recall basic facts, reason solutions, interpret questions, read and translate statements into mathematical equations and make connections between different topics in Mathematics. Hence, by design, many Mathematics tests are multidimensional.

The results of any Mathematics examination have implications for parents, students, teachers, and school administrators: Parents and students are aware that a pass in mathematics is a prerequisite for entry into many universities and for many jobs; teachers are cognizant that

*Corresponding Author. Email: smail.mahdi@cavehill.uwi.edu

ISSN 1727-5512

http://www.cavehill.uwi.edu/fhe/hum/publications/EducationCERJ.htm

judgments about their competence in the class room are 'tied' to the examination results of their students; and, administrators know that their schools are considered "good" or "not good" depending upon the number and quality of passes scored every year at the end of the examination period. It is therefore imperative that the appropriate scoring procedure be used to score multidimensional tests since the number of passes and the ranking of students are likely to vary with each different procedure.

According to Rotou et.*al* (2001) test scores obtained using MIRT procedures are likely to be very different from those obtained using a CTT procedure. Hence, any decision made about a particular individual's success, failure or ranking relative to other students taking the same examination will be dependent upon the particular scoring procedure used. The purpose of this study is to rescore a Mathematics Multiple Choice Test, previously found to be three dimensional, Sealy (2009), using *a* MIRT procedure. Before looking at the MIRT scores, we have organized this paper as follows: firstly, we present the theoretical frame work which includes a summary of the multidimensional model and formulae used to score the mathematics test on each dimension; secondly, selecting the sample, determining the dimensionality and assigning interpretable multidimensional scores are discussed in methodology; thirdly, the method of determining the parameters in the model selected is discussed; fourthly the results and discussions of the CTT scores and MIRT scores are presented; and, finally we conclude with an interpretations of our findings.

Theoretical framework

"Many tests are constructed in such a way that each item presents a problem that can be solved by some characteristic cognitive behaviour that the test intends to measure", see, Kelderman and Rukes (1994). Hence, tests are either one-dimensional or multidimensional. According to Ackerman (1994), in a one-dimensional test, the items measure the same trait or the same multiple of composite traits.

In multidimensional tests, on the other hand, items measure more than one trait; items distinguish between different trait levels; and, examinees have different levels of proficiency on more than one of these traits, see Ackerman (1994) and Reckase and McKinley (1991). In the multidimensional case, Reckase and McKinley (1991) also argued that an "increase in any one – or a combination – of the abilities required for solving the item…" increases the probabilities of answering the item correctly. Ackerman (1994) explains that it is the '.interaction between the characteristics of the item and the latent abilities of the examinees…' that determines the dimensionality of the test and hence the number of latent abilities required for solving the test items.

On the one hand, a CTT total score is the sum of all the items answered correctly and it does not take the dimensionality of the test into consideration; on the other hand, the dimensionality must be known before a test is scored using any MIRT procedures.

Two procedures may be used to determine the MIRT total score. Rotou *et al.* (2001) suggest determining the probability of getting a correct response on each item for each examinee and placing the results in a table which takes the form of an N by K matrix; N being the number of examinees and K, the number of items. The matrix entries are P_{ij} which is the probability that person j gets item i correct. Summing the probabilities in row j, will give the MIRT total score for a particular candidate j. In the second method, the TESTFACT program uses Bayes' (EAP expected aposterori) as explained by Bock and Aitkin (1981) to generate MIRT scores on each of the three dimensions. Two facts are required in calculating MIRT ability scores using Baye's procedure: the number of correct items in the item response vector and which particular items are answered correctly. For example, as stated "If θ_{jk} is the j th ability score, k = 1, 2, 3……..*nfact* for examinee j, j = 1, 2, 3….N, then the factor scores are $E\left(\theta_{ik} \mid x_{i1}, x_{i2}, ...x_{in}\right)$ where x_{ij} is the item j score for the examinee i" (Dutoit, 2003 p. 800). In this paper, only the second method is used because of the amount of calculations required in the Method 1.

For dichotomous data, the model relating the probability of successfully answering an item and the characteristics of the item, which takes guessing in to consideration, was presented by Birnbaum (1968) in the 3PLM as follows:

$$P\left(u_{ij}=1 \mid \theta_j, g_i, b_i\right) = g_i + \frac{1-g_i}{1+e^{\left[-1.7a_i\left(\theta_j - b_i\right)\right]}},$$

and, Embretson and Reise (2000) presented the multidimensional extension of this 3PLM to M3PLM as

$$P\left(u_{ij}=1 \mid \underset{\rightarrow}{\theta}_j, \delta_i, g_i, \underset{\rightarrow}{\alpha}\right) = g_i + \left(1-g_i\right)\frac{e^{\sum_{k=1}^{m}\left(\alpha_{ik}\theta_{jk}+\delta_i\right)}}{1+e^{\sum_{k=1}^{m}\left(\alpha_{ik}\theta_{jk}+\delta_i\right)}}$$

Where θ_{jm} is the trait level of person j on *dimension m;* δ_i, the easiness intercept for item i which is given by $\delta_i = d_i\sqrt{1+\sum_{i=1}^{N}\alpha_{im}^2}$ where d_i is the difficulty of item i, and N is the number of items tested. The term g_i, the pseudo guessing parameter for item i, is given by

$g_i = \dfrac{1}{\# of\ response\, alternatives}$, or it may be calculated with the aid of BILOG, a program used

to estimate parameters in one-dimensional data see, Du Toit (2003).

In this paper, the pseudo guessing parameter was estimated using BILOG.

In the research works of Reckase (1985) and Reckase and McKinley (1991), they developed

formulae that may be used to determine the characteristics of items for a M2PLM.

By extension, for a test with m dimensions, the multidimensional item discrimination may be

written as follows:

$$MDISC_i = \sqrt{(a_{1i}^2 + a_{2i}^2 + ...a_{mi}^2)}\ .$$

(In the M2PLM the parameter m is 2.).

For the multidimensional item difficulty Reckase (1985) defined discrimination in terms of D,

"…the distance to the point of steepest slope in a direction from the origin…" and α, "the angle

needed to describe that direction" as

$$D_i = -\frac{d_i}{MDISC_i}$$

and

$$\alpha_j = \cos^{-1}\left(\frac{a_{1j}}{MDISC}\right).$$

When the characteristics of the items are known, the probability of answering an item correctly in the multidimensional space may be determined.

Using these formulae the multidimensional item discrimination and difficulty were substituted into the M3PLM and the probability of a correct response was determined.

Methodology:
An anonymous proportionate stratified random sample of 3000 Mathematics item response vectors was extracted from the CSEC type Mathematics multiple choice examination. The examination consisted of 60 items and each item presented four response alternatives to each candidate. The result for each item was dichotomized: a "one" was recorded for a response and a "zero" for an incorrect response. Using a Classical Test Theory's total score, the 3000 data values were ranked in descending order. Students who scored 57% or greater were identified as "high scoring students", and all high scoring students were rescored using MIRT; an MIRT score was recorded on each of the three dimensions for all students. Also recorded, is an examinee's MIRT total score found by adding the scores on each dimension. The total score was converted to a T-score for ease of interpretation. The MIRT and CTT scores of the first 30 students and the last 30 students in the ranked high scoring group of 1656 examinees are reported in this, paper.
Each total score was converted to a grade based on the following chart taken from (UWI handbook 2007-2008) which is available online:

Table 1: Grade equivalents.

A+	86-100	B	60-62	C	47-49
A	70-85	B-	57-59	D+	43-46
A-	67-69	C+	53-56	D	40-42
B+	63-66	C	50-52	F	0-39

We assumed that the grades above closely approximate those of the C.X.C examining body.
Hambleton *et al*. (1991), Embretson and Rise (2000) and Bock and Schilling (2003) argued that perfect scores –all items answered correctly or all items answered incorrectly affect IRT procedures. Consequently, all examinees meeting either of these two conditions were removed from the sample before any analysis was done.

The data set was assumed to fit the multidimensional extension of the 3 parameter logistic model. However, before investigating the dimensionality of the data sets, the lower pseudo guessing parameter was estimated to be 0.188 using the BILOG program on the entire sample of 3000 values. Embretson and Rise (2007) pointed out that difficulties sometimes occurred when unique parameters were estimated and they recommend using a common guessing parameter if this happens. That suggestion was followed with this sample and the value 0.188 represents the common guessing parameter.

The pseudo guessing parameter was inputted into TESTFACT which was run in its exploratory mode to *assist* in determining the dimensionality of the item response data set for the high scoring group.

Wood (2003) described the chi square statistic used to determine the dimensionality of the data set as follows:

$\chi^2 = 2\sum_{l=1}^{2^n} r_l \ln \frac{r_l}{N\overline{P_l}}$, with the number of degrees of freedom given as:

Df = $2^n - 1 - n(m+1) + \frac{m(m-1)}{2}$ where 2^n is the number of distinct response patterns and r_l is the frequency of the pattern l. $\overline{P_l}$ is calculated from the maximum likelihood estimates.

$N = \sum r_l$ is the number of cases.

n is the number of items in the test and m is the number of factors.

The TESTFACT program replaced 2^n, in the formula for the number of degrees of freedom, by N_R, the number of cases in the sample, when the data was entered as individual observations. Also, the chi-squared value given in the output was calculated by a slightly different formula:

$$\chi^2 = \sum_{j=1}^{N_R} W_j \ln \frac{W_j}{W_T \times p_j}$$

Where, W_j is the sum of weights for pattern j, W_T is the total sum of weights, and p_j is the marginal probability for pattern j, see, Wood (2003)

The hypothesis tested was H_o: an m factor model fits the data versus the alternative H_1: an m+1 factor model fits the data, and the resultant test statistic was $\chi^2_{df_m - df_{m+1}}$

For a specified level of significance, α, we manually changed the parameter m incrementally and the full information analysis was repeated until $\chi^2_{df_m - df_{m+1}} > \chi^2_{df_m - df_{m+1}, \alpha}$ or $\chi^2_{df_m - df_{m+1}} < 0$ was reached, which ever occurred first. In addition to this chi squared test, both the 2D M3PLM and the 3D M3PLM were fitted to the data set and the model which provided the better fit was selected.

The test constructors subjectively assigned three profile dimensions to the test in the following ratio: P1:P2:P3: 18:24:18. Sealy (2009) showed empirically, using the principles of MIRT, that the data set consisting of the item response vectors of the high scoring group *only* was three dimensional. Consequently, the high scoring group was rescored using the3D, M3PLM and the results were compared to the scores generated using CTT total score.

For easy interpretation the MIRT z scores were converted into T-scores as follows: T-score = 10z +50 where 10 is the standard deviation, 50 is the mean and z is the normalized score, Nitko (1996). For instance, if T = 40 then z = -1 is the corresponding standardized score. So, the T-score values have the appeal of the CTT total score while retaining their relationship with the standardized z scale.

Parameter estimation

TESTFACT was used to estimate the trait levels and item parameters in this paper. The program uses marginal maximum likelihood (MML) to determine the estimates; and, all parameters are determined to minimize errors.

The probability of observing a particular response pattern in an *n-item* test conditioned on ability θ_j is modeled using a likelihood equation. In this paper the 3PLM was used.

Consider the response pattern of item scores for person j:

$\underset{\to}{x}_i = \left[x_1, x_2 .... x_n\right]$.

Since $\underset{\to}{X}_{i1}$ is a random variable with a Bernoulli distribution, from the principle of local independence, it follows that

$$P\left(\underset{\to}{X}_i = \underset{\to}{x}_i \mid \theta_j\right) = P\left(\underset{\to}{x}_i \mid \theta_j\right) = \prod_{i=1}^{n}\left[\phi_i\left(\theta_j\right)\right]^{x_i}\left[1-\phi_i\left(\theta_j\right)\right]^{1-x_i}, \qquad (1)$$

where, $\phi_i\left(\theta_j\right) = P\left(x_i = 1 \mid \theta_j\right) = P_i$.

In any population a number of students will have the same response pattern. Hence using the frequencies of the response patterns, the trait level distribution may be specified from the empirical data: each different response pattern corresponded with a different trait level.

The marginal probability $P(\underset{\sim}{x})$, which is the probability of a particular response pattern for person with unknown ability, according to Embretson and Reise (2000), is given as the sum of the probabilities over the discrete trait levels. If $q = \{1 \ to \ Q\}$ are quadrature points and each quadrature point corresponds with a given latent trait then for the 3PLM

$$P\left(\underset{\sim}{x}_i \mid a_i, d, c_i\right) = \sum_{q=1}^{Q} P\left(\underset{\sim}{x}_i \mid \theta_q, a_i, d, c_i\right) P\left(\theta_q\right) \tag{2}$$

a represents the item's discrimination, d, its difficulty, c its pseudo guessing parameter and θ the latent trait.

The latent trait is considered to be a continuous variable. Taking this into consideration equations 1, for an m-dimensional model, may be written as follows:

$$P\left(\underset{\sim}{x}_i \mid \theta_j\right) = \int_{-\infty}^{\infty} \cdot \cdot \int_{-\infty}^{\infty} \prod_{i=1}^{n} P\left(\underset{\sim}{x}_i \mid \theta_j\right) g\left(\theta\right) d\theta_1 \ldots \ d\theta_m$$

$$= \int_{-\infty}^{\infty} \cdot \cdot \int_{-\infty}^{\infty} \prod_{i=1}^{n} \left[\phi_i\left(\theta\right)\right]^{x_{ij}} \left[1 - \phi_i\left(\theta\right)\right]^{1-x_{ij}} g\left(\theta\right) d\theta_1 \ldots \ d\theta_m \tag{3}$$

$P\left(\underset{\sim}{x}_i \mid \theta\right)$ is the probability of observing a response pattern given some latent trait. And, $g\left(\theta\right)$ is the probability of observing the empirically determined distribution. The integral is evaluated using numerical integration.

In Bock *et.al* (1988) the expression in (1) is symbolized as follows:

$$\overline{P_l} = P\left(\underset{\sim}{x}_i \mid \theta_j\right) = \int_{\theta} L_i\left(\theta\right) g\left(\theta\right) d\theta = \sum_{q_m=1}^{Q} \ldots \sum_{q_2=1}^{Q} \sum_{q_1=1}^{Q} L_i\left(X\right) A\left(X_{q1}\right) \ldots . A\left(X_{qm}\right). \tag{4}$$

Where X is a multidimensional quadrature point, and A(X) is the weight corresponding to a given quadrature point.

Let r_l be the frequency of a given response pattern $\underset{\to}{x}_i$ for an n item test in a sample of N examinees. If there are s distinct patterns then $l = 1\ to\, s \le \left(N, 2^n\right)$ and $\sum_{l=1}^{s} r_l = N$. Hence, r_l has a multinomial distribution with parameter N. And, the probability of the sample data likelihood is given as follows:

$$L_m = P\left(X\right) = P\left(r_1, r_2.....r_s; N\,\overline{P}_1\,\overline{P}_2.....\overline{P}_s\right) = \frac{N!}{r_1!..r_s!}\overline{P}_1^{r_1}\,\overline{P}_2^{r_2}.....\overline{P}_s^{r_s}.$$

Taking logs of both sides we get the following:

$$\ln L_M = Ln\,P\left(X\right) = \ln\left(\frac{N!}{r_1!..r_s!}\right) + \sum_{l=1}^{s} r_l \ln \overline{P}_l \tag{5}$$

This shows that the log likelihood is a function of the marginal probability weighted by the frequency of the given response pattern. To determine estimates of the item parameters we look for the values that maximize the data log likelihood. That is $\ln L_m$ is differentiated and the result is set equal to zero. Bock et.al (1988) differentiated $\ln L_m$ with reference to some general item parameter v_i and got the following results:

$$\frac{\partial \ln L_m}{\partial v_j} = \sum_{l}^{s} \frac{r_l}{\overline{P}_l} \int_{\theta} \left(\frac{x - \phi}{\phi\left(1 - \phi\right)}\right) L_l \frac{\partial \phi}{\partial v_j} g\left(\theta_j\right) d\left(\theta_j\right) = \int_{\theta} \frac{\overline{r}_j - \overline{N}\phi}{\phi\left[1 - \phi\right]} \cdot \frac{\partial \phi}{\partial v_j} g\left(\theta_j\right) d\left(\theta_j\right). \tag{6}$$

where

$$\overline{r}_j = \sum_{l=1}^{s} \frac{r_l x_{lj} L_l\left(\theta_j\right)}{\overline{P}_l} \tag{7}$$

And

$$\overline{N} = \sum_{l=1}^{s} \frac{r_l L_l\left(\theta_j\right)}{\overline{P}_l} \tag{8}$$

According to Bock et.al (1988), the integral in 3 which corresponds with the maximization procedure in the EM algorithm may be calculated by numerical integration methods, and 4 and 5 represent expectation steps. In the expectation stage according to Embretson and Reise (2000), the expected number of persons at each trait level (represented by quadrature points) is computed as well as the number students passing each item.

With reference to the 3PLM these frequencies are substituted into the estimation equations 6 below, and the values of the parameters that maximize these likelihood equations:

$$\sum_{k=1}^{q}\left(\frac{\bar{r}_{ik} - \overline{N}_{ik} P_i\left(X_k\right)}{P_i(X_k)[1-P_i(X_k)}\right)\frac{\partial P_i(X_k)}{\partial\begin{bmatrix} c_i \\ a_i \\ g_i \end{bmatrix}} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{9}$$

are found in Du Toit (2003) where,

$\bar{r}_{ik} = \sum_{l}^{s} r_l x_{li} P(x_l \mid X_k) \frac{A(X_k)}{P_{xl}}$, such that x_{li} is the score 0 or 1 for item i in pattern l, and

$\overline{N}_k = \sum_{l}^{s} r_l P(x_l \mid X_k) \frac{A(X_k)}{P_{xl}}$.

If convergence is not reached a second expectation procedure is used with the new estimates and a second maximization procedure is also used. When convergence is achieved a Newton-Gauss procedure is used to get the final estimates Embretson and Reise (2000).

Results and discussion

Table 2: CTT and MIRT scores for the first 30 examinees

SCORES OF THE FIRST 30 EXAMINEES

		CTT					MIRT		
Person	P1	P2	P3	TOTAL	Dim1	Dim2	Dim3	ztotal	T-score
1	18	24	17	59	1.225	0.772	1.899	3.896	88.96
2	18	23	18	59	1.829	0.213	1.57	3.612	86.12
3	18	23	18	59	1.689	0.787	1.411	3.887	88.87
4	18	24	17	59	1.259	0.237	1.842	3.338	83.38
5	18	23	18	59	1.689	0.787	1.411	3.887	88.87
6	18	24	17	59	1.747	0.59	1.672	4.009	90.09
7	18	23	18	59	1.237	0.85	1.672	3.759	87.59
8	18	24	17	59	1.747	0.59	1.672	4.009	90.09
9	18	24	17	59	1.225	0.772	1.899	3.896	88.96
10	18	24	17	59	1.747	0.59	1.672	4.009	90.09
11	18	24	17	59	1.747	0.59	1.672	4.009	90.09
12	18	23	17	58	1.69	0.707	1.216	3.613	86.13
13	18	23	17	58	1.245	0.215	1.77	3.23	82.3
14	16	24	18	58	1.708	0.644	1.514	3.866	88.66
15	17	23	18	58	1.36	0.714	1.254	3.328	83.28
16	17	24	17	58	0.897	1.004	1.545	3.446	84.46
17	18	23	17	58	1.729	0.909	1.434	4.072	90.72
18	18	23	17	58	1.797	0.101	0.962	2.86	78.6
19	17	23	18	58	1.599	0.636	1.48	3.715	87.15
20	17	24	17	58	1.619	0.456	1.229	3.304	83.04
21	18	22	18	58	1.343	0.753	1.286	3.382	83.82
22	17	23	18	58	1.639	0.848	1.202	3.689	86.89
23	18	24	16	58	1.248	-0.044	1.61	2.814	78.14
24	18	23	17	58	1.171	0.972	1.568	3.711	87.11
25	18	23	17	58	1.171	0.972	1.568	3.711	87.11
26	18	22	18	58	1.809	0.52	1.325	3.654	86.54
27	18	23	17	58	1.971	0.797	1.343	4.111	91.11
28	18	23	17	58	1.773	0.252	0.5	2.525	75.25
29	18	22	18	58	1.658	-0.49	1.668	2.836	78.36
30	18	22	18	58	1.724	-0.12	1.668	3.272	82.72

Table 3: CTT and MIRT scores for the bottom 30 examinees

SCORES OF THE LAST 30 EXAMINEES

	CTT				MIRT				
person	**P1**	**P2**	**P3**	**total**	**Dim1**	**Dim2**	**Dim3**	**Ztotal**	**TSCORE**
1627	10	16	8	34	-0.781	-0.885	-0.848	-2.514	24.86
1628	11	16	7	34	-1.536	-0.15	-0.311	-1.997	30.03
1629	9	16	9	34	-1.144	-0.189	-0.67	-2.003	29.97
1630	11	15	8	34	-0.529	-0.407	-0.933	-1.869	31.31
1631	10	14	10	34	-0.031	-0.719	-1.58	-2.33	26.7
1632	11	14	9	34	-0.584	-0.656	-1.012	-2.252	27.48
1633	13	9	12	34	-2.034	-0.988	0.434	-2.588	24.12
1634	12	14	8	34	-0.985	-1.058	-0.18	-2.223	27.77
1635	10	14	10	34	-0.224	-2.583	0.135	-2.672	23.28
1636	10	15	9	34	0.305	-2.644	-0.484	-2.823	21.77
1637	10	16	8	34	0.499	-2.025	-0.652	-2.178	28.22
1638	9	12	13	34	0.327	-2.125	-1.141	-2.939	20.61
1639	13	14	7	34	-1.246	-2.075	0.419	-2.902	20.98
1640	8	17	9	34	-0.251	-2.684	0.207	-2.728	22.72
1641	10	15	9	34	-0.171	-2.411	0.305	-2.277	27.23
1642	12	15	7	34	-1.111	-2.222	0.548	-2.785	22.15
1643	12	13	9	34	-0.708	-0.946	-0.641	-2.295	27.05
1644	11	13	10	34	-1.213	-0.801	-0.2	-2.214	27.86
1645	10	15	9	34	-0.782	-0.051	-1.04	-1.873	31.27
1646	11	12	11	34	-1.067	0.222	-1.026	-1.871	31.29
1647	**11**	**17**	**7**	**35**	**-0.949**	**-0.752**	**-0.477**	**-2.178**	**28.22**
1648	8	15	11	34	-0.138	-0.55	-1.559	-2.247	27.53
1649	12	16	6	34	-1.166	-0.088	-0.731	-1.985	30.15
1650	12	12	10	34	-0.967	-0.645	-0.358	-1.97	30.3
1651	13	12	9	34	-0.274	0.002	-1.602	-1.874	31.26
1652	12	12	10	34	-1.779	0.016	-0.599	-2.362	26.38
1653	10	13	11	34	-0.212	0.151	-2.007	-2.068	29.32
1654	10	14	10	34	-0.946	0.445	-1.39	-1.891	31.09
1655	11	14	9	34	-0.626	-0.243	-1.062	-1.931	30.69
1656	9	15	10	34	-0.379	-0.647	-1.26	-2.286	27.14

From **table 1** the range of the CTT total marks = 59-58 = 1 and the range of the MIRT total marks = 91.11 – 75.25 = 15.86 for the first 30 students. In **Table 2** the range of the CTT total marks = 34-34 = 0 and the range of the MIRT total mark = 31.31 – 20.61= 10.70. If students are ranked according to total test scores, under CTT there would only be two positions in the top group: 11 persons would be in first position and 19 persons would be in second position. In the bottom group there would also only be two positions: 29 students scored the same mark and only one student got a different mark using CTT scoring methods. However under MIRT scoring procedures, each student in the two groups identified was assigned a unique score and hence a unique rank order position.

Conclusion

If the examination results are based on a total test score, students who get a high total score using the CTT analysis may not get a similar high score with the MIRT procedure. In this study, many students were given a failing grade when the MIRT methodology was employed and the total MIRT score was the only criterion used to estimate the student's ability; yet, these same students got a passing grade using the CTT procedure. Clearly, if the total test score is being used to evaluate a student's overall performance the *results* will be dependent upon which methodology is used - CTT or MIRT.

However, a total test score does not give an accurate impression of a person's mathematical ability on a multidimensional test. Modeling the data multi-dimensionally allows the analyst to make inferences about an examinee on *each* of the latent traits needed to describe the entire latent space. In this paper three constructs were assigned to the test by test constructors: Recall, reasoning and comprehension. The dimensionality of the item response data set was also found to be three. Therefore, it is our view that all candidate should be accessed on each of the three dimension empirically determined; consider a student, for example, who gets a failing grade. This student may have demonstrated competence in reasoning ability – one of the three dimensions measured by the test. If this quality is important to a potential employer or a graduate advisor it will go unnoticed since a failed subject may not be considered by either person. According to Gosz and Walker (2002), a Classical Test Theory total test score on a multidimensional test could lead to an inaccurate estimation of a person's ability. That point has been corroborated in our study. Hence, we believe that a multidimensional test should be modeled using a multidimensional procedure; and, that special attention needs to be paid to the scores on each dimension and what information is reported, whether the overall performance is a passing or a failing grade.

References

Ackerman, T. A. 1994. Using multidimensional item response theory to understand what Items and tests are measuring. *Applied Measurement in Education* 7(4): 255-278.

Birnbaum, A. 1968. Some latent trait models. In *Statistical theories of mental test scores, Frederic Lord and Melvin Novic, 397-422*. Reading. MA:Addison-Wesley

Bock, R.D., R. Gibbons & E Muraki.1988. Full information item factor analysis. *Applied Psychological Measurement* 12: 261-280.

Bock, R.D. and M.Aitkin.1981. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* 46: 443-459.

duToit, Mathilda, ed. 2003. *IRT from SSI: BILOG-MG, MULTILOG, PARASCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Embretson, Susan, and Steven Reise. 2000. *Item response theory for psychologists*. NJ: Lawrence Erlbaum Associates.

Hambleton, Ronald., H Swaminathan, and H. Jane Rogers. 1991. *Fundamentals of item response theory*. Newbury Park, CA: Sage

Nitko J. Anthony. 1996. *Educational assessment of students*. NJ: Prentice Hall Inc.

Reckase, M. D., and R L. McKinley. 1983. The definition of difficulty and discrimination for multidimensional item response theory models. Paper presented at the Annual Meeting of the American Education Research Association, April 11-15. Montreal, Quebec.

Reckase, M.D., and R.L. McKinley. 1991. The discrimination power of items that measure more than one dimension. *Applied Psychological Measurement 15(4): 361-373.*

Reckase, M.D.1997.The past and future of multidimensional item response theory. *Applied Psychological Measurement* 21: 25-36.

Reckase, M .D.1985. The difficulty of test items that measure more than one ability. *Applied Psychological Measurement* 9: 401- 412.

Rotou, Suranian, Patricia B.Elmore, and Todd C. Headrick. 2001. Comparison between classical true score theory and multidimensional item response theory. Paper presented at the Annual Meeting of the American Educational Research Association, April 10 – 14. Seattle, WA.

Rukes, Carl. P.M. and Henk Kelderman. 1994. Loglinear multidimensional irt models for polytomously scored items. *Psychometrika* 59(2): 149 – 176.

Sealy, P. 2009. Assessing the construct validity of CXC mathematics multiple choice examination using MIRT. M.Phil. thesis., Univ. of the West Indies.

Bock D. and Schilling S. G. 2003. IRT based item factor analysis. In *IRT from SSI:* BILOG-*MG, MULTILOG, PARASCALE, TESTFACT, Mathilda du Toit, ed. 585-591.* Lincolnwood. IL: Scientific Software International.

Walker, Cindy M., and Natasha S. Beretvas. 2000. Using multidimensional versus unidimensional ability estimates to determine student proficiency in mathematics. Paper presented at the Meeting of the American Educational Research Association April, 24-28, in New Orleans, LA.

Walker, Cindy M., and Jacqueline K. Gosz. 2002. An Empirical comparison of multidimensional item response data using TESTFACT and NOHARM. Paper presented at the Annual Meeting of the National Council for Measurement in Education (NCME) April 2-4 in New Orleans, Louisiana.

Wood Robert 2003.Statistical test of the number of factors. In *IRT from SSI:* BILOG-*MG, MULTILOG, PARASCALE, TESTFACT, Mathilda du Toit, ed. 585-591.* Lincolnwood. IL: Scientific Software International.

Wood Robert 2003. TESTFACT examples. In *IRT from SSI:* BILOG-*MG, MULTILOG, PARASCALE, TESTFACT, Mathilda du Toit, ed. 585-591.* Lincolnwood. IL: Scientific Software International.