International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

International Journal of Sciences: Basic and Applied Research

ISSN 2307-4531
(Print & Online)

Published by:
GSSRR

Visit: www.gssrr.org

A Comparative Study on Classical Test Theory and Rasch Model in *Comprehensive Mental Ability Test (CMAT)*: A Pilot Test

Josephine P. Manapsal*

Department of Education, Division of Cavite-Francisco Osorio National High School Trece Martires City, Cavite 4109, Philippines, College of Arts and Sciences- Cavite State University, Indang Cavite 4109 Philippines
Email: prinsesamj@yahoo.com /josmanapsal@yahoo.com

Abstract

This study is a comparative analysis of Classical Test Theory (CTT) and the Rasch Model in the pilot testing of Comprehensive Mental Ability Test (CMAT). It investigated whether bad items in CTT would also be bad or unfit items when the Rasch model is applied. Thirty-three Master's level students in one of the state universities in Manila were the participants of the study. Sixty minutes were given for them to answer the 60 items of CMAT. Using CTT, the reliability test result showed that the computed KR#20 is .76 while using Rasch, the computed person ability reliability is .73.Out of 60 items, 5 items were found to be bad items using CTT and 17 items were found to be misfit using Rasch Analysis. This result showed that using the Rasch Model, it becomes stricter than CTT. Also, there are items discarded in CTT but not in the Rasch model, and there are items found to be unfit in the Rasch model but they are good items in the CTT. Thus, each particular model has their own specific parameter, whether which one will offer a better outcome for test construction and development, the answer will rely on the parameter of the tests, orientation of the test developer and the purpose of the test.

Keywords: Item difficulty; item discrimination; IRT; item calibration; misfit items; unidimensionality; overfit; *infit; logits.*

----

* Corresponding author.

1. Introduction

A pilot test for the *Comprehensive Mental Ability Test (CMAT)* was conducted in order to evaluate constructed items that will actually measure an individual's potential to learn and to adapt in academic environment. The CMAT is an attempt to investigate a person's cognitive functions. It is designed to measure one's comprehensive verbal and non verbal mental ability. The CMAT measures the person's cognitive abilities and to determine learning difficulties in a particular area. It is a comprehensive and intensive battery of tests that has various subtests. It measures one's specific ability and diagnoses one' difficulty in a particular area. Six (6) subscales of this test measures the six cognitive faculty: 1. Language Structure; 2. Verbal Analogy; 3. Judgment and Comprehension; 4.Arithmetic; 5. Reading Comprehension; 6. Perceptual Acuity. These are anchored on the basic mental abilities of a student needed in order to survive in an academic setting.

There are two widely perceived measurement models to determine the items' validity and reliability. One is a psychometrically-sound traditional measurement theory known as the CTT, the other one is known as the Modern Trait Theory which is widely known as the Item Response Theory IRT (others preferred to call it the latent trait theory) which gained popularity because of its promise to provide greater precision and control in measurement [7]. Item Response Theory contains a large family of models, one of which is the Rasch model (1960) which became popular in the 70's. The CTT, on the other hand, became popular ever since testing was introduced because it is simple and practical to use. It has no complex computations that would require a strong statistical ability and does not require the use of computer software programs.

Georg Rasch developed a mathematical model for constructing measures based on a probabilistic relation between an item's difficulty and a person's ability [22]. The Rasch model is the simplest among the IRT models because it uses only one parameter which is an *item difficulty* to specify item characteristics [6]. This model is described as one-parameter logistic (IPL) latent trait model because it only contains one item parameter [4]-[10]. introduced the idea of a latent trait or ability and differentiated this construct from observed test score. The unobservable (or latent) construct being measured by the questionnaire is usually expressed in the formula of theta or θ [19].

In the Rasch family of models, there are four models: 1. dichotomous model, 2. polytomous or rating scale model, 3. partial credit model and 4.many-facets Rasch Model. The dichotomous model is used in this study, because the data used have only two parameters, either the person gets the correct or incorrect response, or passes or fails the item.

The reason for applying both the CTT and One Parameter Logistic Model or Rasch model is twofold: 1. to establish good/fit and bad/unfit items that will either be retained or eliminated and whether the same items will be retained or eliminated in the CMAT and; 2. to identify the strength and limitations of CTT and IRT in test construction and development.

In test validation, still the widely used approach in the country is the Classical Test Theory that centers in the two item statistics: *item difficulty* (the proportion of sample population who got the correct answer) and *item* *discrimination* (the correlation between passing the item and some measure of ability). Despite its popularity, CTT has a number of shortcomings. First, CTT is *person-dependent*; that is, the result cannot generalize item difficulty and discrimination statistics across groups and this limitation reflects a problem of circular dependence and, as such, the characteristics of items cannot be separated from the characteristics of the examinees [7]. Hence, there is a need to change or set new norms to adapt to another group of population depending on the use and function of the test. Second, CTT is *item- dependent*; that is the person ability statistic (the observed score for a given set of items) is a function of the difficulty of the sample of item administered [7]. The score and the norm are dependent on the performance of the sample group included in the test. Thus, item difficulty varies depending on the sample of test takers who take the specific test. In contrast, Rash model calibration process theoretically makes the ability statistic item-free and item difficulties examinee free [7]. The item calibration in test item difficulty is independent of the person used for the calibration. It means that when the instrument is administered, it will give the same results regardless of who takes the test. Third, CTT works on the assumption that the measurement error variance is the same scores of all persons to whom the instrument was administered [7] whereas, in the Rasch model, standard errors for individual ability estimates are computed instead of just a single estimate for all the test takers. The estimates of item difficulty remain constant from one sample of persons to another; at least as long as the data fit the model [6].

The CTT on the other hand, has advantages over the Rasch model. The CTT can hold a large sampling population to determine its reliability. In fact, as the sample population gets bigger, the CTT works better. In contrast, Rasch model is limited to a very large scale testing. Second, CTT is a lot simpler as compared to Rasch analysis, where manual calculations for item indices and ability estimates involve a tedious method. To verify the weakness of CTT, this study is conducted.

2. Methodology

2.1 Participants

The Comprehensive Mental Ability Test (CMAT) is composed of 60 items and administered as pilot testing to thirty-three (33) volunteer graduate students (*N* = 33; 75.8 % women, 24.4 % men) in one of the state universities in Manila. They were given 60 minutes to answer the test.

A single answer sheet was provided for all types of sub-tests. The participants were asked not to leave any of the personal data asked for. They were then asked to read silently the directions on the cover of the booklet while the examiner read them aloud and clear. Questions were solicited for clarifications. Participants were then asked to answer each sub tests as fast and as accurately as they could. Each sub-tests is administered in 10 minutes.

2.2 Instrument

CMAT is designed to measure six areas of mental ability.

1. Language Structure. It is composed 10 items that incorporates grammar rules in the Subject-Verb

Agreement. It measures one's ability to distinguish the correct and incorrect grammatical syntax of the English language.

2. Verbal Analogy . It consists of 10 items of general information arranged in word-pairs. It measures a person's comprehension, accuracy and mental speed in determining the nature of relationship between the word-pair.
3. Judgment and Comprehension. This segment has two divisions; one is composed of 5 items that measure one's sense of logic in the given premises. The second part measures one's ability or inability to make judgment or conclusion from the given assumptions. This is to measures one's learning aptitude in distinguishing valid and invalid arguments.

Table 1: Table of Specifications for CMAT

Domain	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation
LSLA	1,2,3,4,5, 6, 7, 8, 9, 10					
VALA				11,12,13,14 ,15,16,17,18, 19, 20		
JCLA						21,22,23,24 25,26,27,28 29,30
ALA			31,32, 33, 34,35, 36, 37,38,39,40			
R & C		41,42,43, 44,45,46,47 48, 49, 50				
PA					50, 52,53 54,55,56, 57,58,59,60	

Legend:

LSLA - *Language Structure Learning Aptitude*;

VALA - *Verbal Analogy Learning Aptitude*;

JCLA - *Judgment and Comprehension Learning Aptitude*;

ALA - *Arithmetical Learning Aptitude;*

R & C- *Reading and Comprehension*;

PA - Perceptual Acuity

4. Mathematical Ability. This sub test is composed of 10 variations of numerical problems required to be solved with accuracy and mental speed. This measures the person's quantitative reasoning facility such as dealing accurately and efficiently with numerical figures.
5. Reading and Comprehension. This is composed of 10 items with 3 given passages. It measures how the person conceptualizes; and deduces concepts in the given passages. It measures the person's ability to integrate concepts and good judgment into a meaningful and comprehensive relation.

6. Perceptual Acuity. It consists of 10 items that measure one's ability or learning aptitude to perceive with non-verbal, non-numerical presentation of visio-spatial problems. This requires one's mental speed and accuracy in determining analogical and other types of relationships. Table 1 shows the table of specifications for CMAT.

2.3 Validity

CMAT test items underwent a *content validation* from the experts in tests construction and have published articles related to testing, they are members of the faculty of the two graduate schools in Manila. Aside from *content validation,* CTT approach was utilized. Using CTT, item analysis, item discrimination and item difficulty was computed to identify good and bad items.

Item analysis in CTT is the procedure to identify the difficulty of the item and the discrimination of the item. Item difficulty refers to the level or degree on how easy or difficult is the item using the difficulty index. Whereas, measures of Item discrimination indicates how adequately an item separates or discriminates between high scorers and low scorers of an entire test [2].

Item difficulty is computed by using the following formula

$$P = \frac{P_H + P_L}{N}$$

whereas, item discrimination is computed by

$$D = \frac{D_H - D_L}{N_1 \quad \text{or} \quad N/_2}$$

Another means of validation of procedure used was the use of the Rasch Model with the use of Statistical software named, WINSTEP. In the Rasch Model, item calibration is a preferred term instead of *item analysis*. In item calibration, the preferred term is *"fit"* or *"unfit"* or "misfit" instead of labeling items as *"good*" or "*bad"* items.

In the Rasch model, item difficulty estimates are expressed in *logits,* in which a *logit* value of 0 is arbitrarily set as the average or the mean, the positive *logit* estimates indicates that the items are progressively more difficult. On the other hand, *person ability* is estimated in relation to the item difficulty estimates where the more negative the value, the lower the ability of the person to perform well in the test [22].

2.4 Reliability

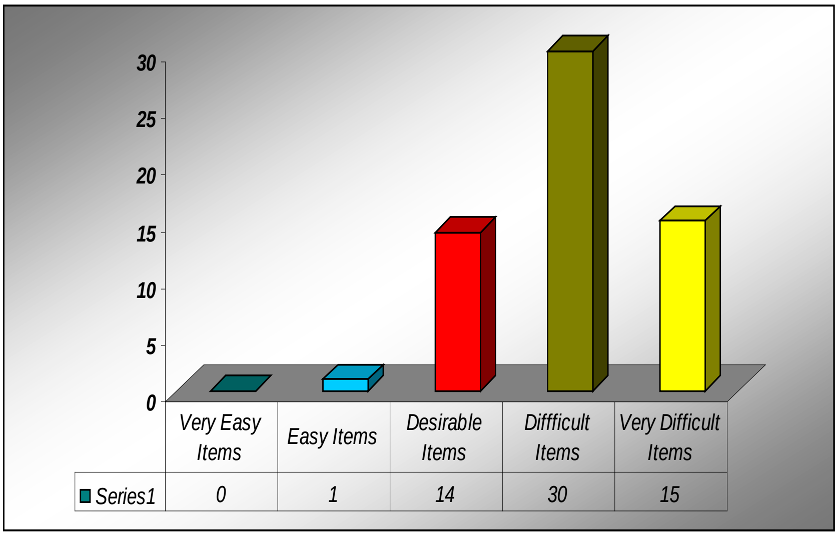
In CTT, reliability coefficient refers to the extent to which the test is likely to produce consistent scores that reflects the two characteristics of the test: 1.) the intercorrelations among the items; 2.) the length of the test. The intercorrelations among the items refer to the greater the number of positive relationships, and the stronger those relationships are, the greater the reliability while the length of the test is correlated with reliability. In CTT, a test with more items will have a higher reliability; hence item banking is commonly the practice of item developers.

Reliability of a measurement concept plays the most crucial part in classical theory because this reflects the accuracy with which a group of individuals can be ranked or ordered on the basis of test performance and establishing norms [6]. In contrast, "In the Rasch model the concept of reliability plays a subordinate part, because this measurement model is oriented toward estimation of individual ability, rather than comparisons of [6].

Using CTT, the reliability result showed that the computed KR#20 is .76 (Mean = 32.96, SD =6.80; Variance= 46.22), while using Rasch model, the computed person ability reliability is .73.The estimate errors associated are small indicating that the data fit well the expected ability and test difficulty.

3. Results and Discussion

Of the 33 graduate students who participated in the study, though relatively small for analysis using CTT, 0.27 percent of the *upper group* and *lower group* were used to compute for item difficulty and item discrimination. The upper group (Upper Limit) refers to the 0.27 percent of the test takers who got the *correct* answers, while the lower group (Lower Limit) refers to the 0.27 percent of the test takers who got the *wrong* answer. The middle portion of the group is not part of the data analysis.



	Very Easy Items	Easy Items	Desirable Items	Diffficult Items	Very Difficult Items
■ Series1	0	1	14	30	15

Figure 1: Item Difficulty Analysis of CMAT 60 items

Figure 1 shows that Item Difficulty Analysis in CTT reveals that half of the items are found to be difficult. Fifteen (15) out of 60 items are found to be very difficult while fourteen (14) items are found to be good or desirable items. Only one item is easy and none of the items was found to be very easy.

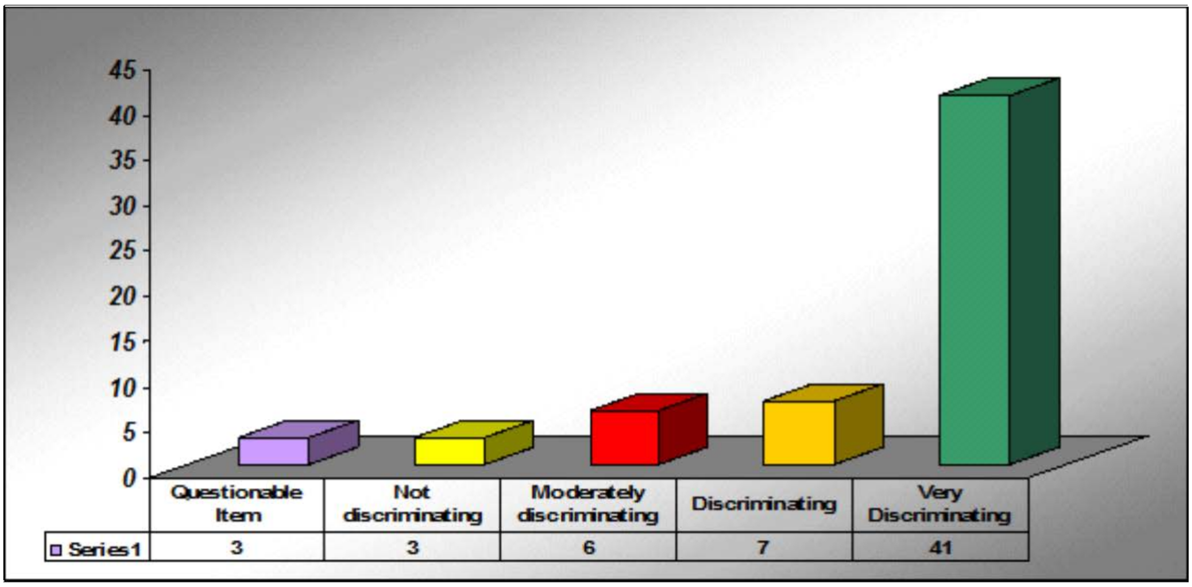


Figure 2: Item Discrimination Analysis of CMAT in 60 items

Figure 2 shows that out of 60 items, forty one (41) items are found to be very *discriminating*. It means that high scorers and low scorers can be discriminated easily. Seven (7) items are found to be discriminating and six (6) items are moderately discriminating. Three (3) items are found to be not discriminating and three (3) items are found to be questionable. Questionable items are neither easy nor difficult to discriminate high scorers from low scorers.

Table 2: The summary of items to be accepted and discarded for the CMAT

				DS			**DF**		
ITEMS	**UL**	**LL**	**DS**	**DESCRIPTION**	**DECISION**	**DF**	**DESCRIPTION**	**DECISION**	**FINAL SELECTION**
1	4	1	0.33	Difficult	Retain	0.28	Reasonably Good	Moderately discriminating	**ACCEPT**
2	3	1	0.22	Difficult	Retain	0.22	Reasonably Good	Moderately discriminating	**ACCEPT**
3	6	4	0.22	Difficult	Retain	0.56	High	Very Discriminating	**ACCEPT**
4	4	1	0.33	Difficult	Retain	0.28	Reasonably Good	Moderately discriminating	**ACCEPT**
5	6	1	0.56	Desirable	Retain	0.39	Satisfactory	Discriminating	**ACCEPT**
6	5	4	0.11	Very Difficult	Revise/ Reject	0.5	High	Very Discriminating	**ACCEPT**
7	4	1	0.33	Difficult	Retain	0.28	Reasonably Good	Moderately discriminating	**ACCEPT**
*8	1	2	-0.11	Very Difficult	Revise/ Reject	0.17	Marginal	Not discriminating	**DISCARD**
9	5	3	0.22	Difficult	Retain	0.44	High	Very Discriminating	**ACCEPT**
10	4	3	0.11	Very Difficult	Revise/ Reject	0.39	Satisfactory	Discriminating	**ACCEPT**
11	5	1	0.44	Desirable	Retain	0.33	Satisfactory	Discriminating	**ACCEPT**

12	7	6	0.11	Very Difficult	Revise/ Reject	0.72	High	Very Discriminating	**ACCEPT**
13	7	3	0.44	Desirable	Retain	0.56	High	Very Discriminating	**ACCEPT**
14	5	3	0.22	Difficult	Retain	0.44	High	Very Discriminating	**ACCEPT**
15	7	7	0	Very Difficult	Revise/ Reject	0.78	High	Very Discriminating	**ACCEPT**
16	5	2	0.33	Difficult	Retain	0.39	Satisfactory	Discriminating	**ACCEPT**
17	8	5	0.33	Difficult	Retain	0.72	High	Very Discriminating	**ACCEPT**
18	9	7	0.22	Difficult	Retain	0.89	High	Very Discriminating	**ACCEPT**
19	6	5	0.11	Very Difficult	Revise/ Reject	0.61	High	Very Discriminating	**ACCEPT**
20	9	9	0	Very Difficult	Revise/ Reject	1	High	Very Discriminating	**ACCEPT**
*21	0	1	-0.11	Very Difficult	Revise/ Reject	0.06	Poor	Questionable Item	**DISCARD**
*22	0	1	-0.11	Very Difficult	Revise/ Reject	0.06	Poor	Questionable Item	**DISCARD**
23	4	4	0	Very Difficult	Revise/ Reject	0.44	High	Very Discriminating	**ACCEPT**
24	5	2	0.33	Difficult	Retain	0.39	Satisfactory	Discriminating	**ACCEPT**
25	4	2	0.22	Difficult	Retain	0.33	Satisfactory	Discriminating	**ACCEPT**
26	9	4	0.56	Desirable	Retain	0.72	High	Very Discriminating	**ACCEPT**
27	8	5	0.33	Difficult	Retain	0.72	High	Very Discriminating	**ACCEPT**
28	7	4	0.33	Difficult	Retain	0.61	High	Very Discriminating	**ACCEPT**
29	7	3	0.44	Desirable	Retain	0.56	High	Very Discriminating	**ACCEPT**
30	4	1	0.33	Difficult	Retain	0.28	Reasonably Good	Moderately discriminating	**ACCEPT**
31	6	1	0.56	Desirable	Retain	0.39	Satisfactory	Discriminating	**ACCEPT**
*32	0	1	-0.11	Very Difficult	Revise/Reject	0.06	Poor	Questionable Item	**DISCARD**
33	3	1	0.22	Difficult	Retain	0.22	Reasonably Good	Moderately discriminating	**ACCEPT**
34	7	6	0.11	Very Difficult	Revise/Reject	0.72	High	Very Discriminating	**ACCEPT**
35	8	5	0.33	Difficult	Retain	0.72	High	Very Discriminating	**ACCEPT**
36	9	4	0.56	Desirable	Retain	0.72	High	Very Discriminating	**ACCEPT**
37	7	4	0.33	Difficult	Retain	0.61	High	Very Discriminating	**ACCEPT**
38	6	5	0.11	Very Difficult	Revise/Reject	0.61	High	Very Discriminating	**ACCEPT**
39	9	6	0.33	Difficult	Retain	0.83	High	Very Discriminating	**ACCEPT**
40	9	4	0.56	Desirable	Retain	0.72	High	Very Discriminating	**ACCEPT**
41	8	4	0.44	Desirable	Retain	0.67	High	Very Discriminating	**ACCEPT**

*42	4	6	-0.22	Very Difficult	Revise/Reject	0.56	High	Very Discriminating	**DISCARD**
43	7	2	0.56	Desirable	Retain	0.5	High	Very Discriminating	**ACCEPT**
44	3	0	0.33	Difficult	Retain	0.17	Marginal	Not discriminating	**ACCEPT**
45	3	0	0.33	Difficult	Retain	0.17	Marginal	Not discriminating	**ACCEPT**
46	7	3	0.44	Desirable	Retain	0.56	High	Very Discriminating	**ACCEPT**
47	8	5	0.33	Difficult	Retain	0.72	High	Very Discriminating	**ACCEPT**
48	8	7	0.11	Very Difficult	Revise/Reject	0.83	High	Very Discriminating	**ACCEPT**
49	8	5	0.33	Difficult	Retain	0.72	High	Very Discriminating	**ACCEPT**
50	9	3	0.67	Easy	Revise	0.67	High	Very Discriminating	**ACCEPT**
51	7	2	0.56	Desirable	Retain	0.5	High	Very Discriminating	**ACCEPT**
52	8	6	0.22	Difficult	Retain	0.78	High	Very Discriminating	**ACCEPT**
53	9	4	0.56	Desirable	Retain	0.72	High	Very Discriminating	**ACCEPT**
54	5	3	0.22	Difficult	Retain	0.44	High	Very Discriminating	**ACCEPT**
55	8	5	0.33	Difficult	Retain	0.72	High	Very Discriminating	**ACCEPT**
56	9	4	0.56	Desirable	Retain	0.72	High	Very Discriminating	**ACCEPT**
57	5	3	0.22	Difficult	Retain	0.44	High	Very Discriminating	**ACCEPT**
58	8	5	0.33	Difficult	Retain	0.72	High	Very Discriminating	**ACCEPT**
59	9	7	0.22	Difficult	Retain	0.89	High	Very Discriminating	**ACCEPT**
60	9	6	0.33	Difficult	Retain	0.83	High	Very Discriminating	**ACCEPT**

**** deleted items***

UL - *Upper Limit* DS - *Item Discrimination*

LL - *Lower Limit* DF -*Item Difficulty*

Table 3: Discrimination Index scale used in CMAT

0.10 and below	Poor	Questionable Item
0.11 to 0.20	Marginal	Not discriminating
0.21 to 0.30	Reasonably Good	Moderately discriminating
0.31 to 0.40	Satisfactory	Discriminating
0.41 above	High	Very Discriminating

Table 3 shows the discrimination index scale used in CMAT. Poor and marginal items have to be deleted.

Table 4: Difficulty Index scale used in CMAT

0.00 to 0.20	Very Difficult	Revise/Reject
0.21 to 0.40	Difficult	Retain
0.41 to 0.60	Desirable	Retain
0.61 to 0.80	Easy	Revise
0.81 to 1.00	Very Easy	Revise/Reject

Table 4 shows the difficulty index used in CMAT. Items that turned out to be very difficult or very easy were either have to revise or reject.

Item Calibration in Rasch Model

The term, *item calibration* is preferred more than *item analysis* when using the Rasch Model. Both terms may mean the same but the process of measurement for each is different. *Item calibration* is a procedure of estimating a person ability expressed in item difficulty by converting raw scores to logits on an objective measurement scale. Logit is a unit of measurement that results when the Rasch model is used to transform raw scores obtained from ordinal data to log odd ratios on a common interval scale in which the value of 0.0 is routinely allocated to the means of the item difficulty estimates [22]

The responses of 33 graduate students to a single administration of the 60 item CMAT is computed using WINSTEPS assisted by a statistician. Technically, the WINSTEPS, the computer software for Rasch measurement can construct measures from simple rectangular data sets, usually of persons and items, with up to 1,000,000 cases and 10,000 items using Joint Maximum Likelihood Estimation JMLE by [22].

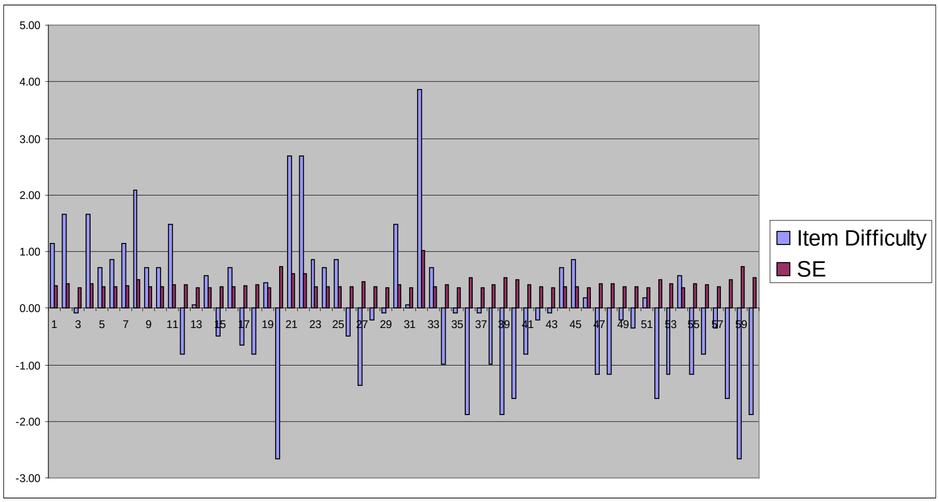


Figure 3: Log Estimates of Item Difficulty and SE

Figure 3 shows the log estimates of item difficulty of 60 items with their corresponding standard errors. The log estimates functions for each item shows small standard errors. The error estimates range from .36 to 1.02. It means that the items are relatively stable, hence, the instrument is considered to have its learning ability to accurately measure.

Table 5: Item Estimates for all students (n=33)

Item No.	Difficulty Estimate	Error Estimate	Infit Mean Square	Outfit Mean Square
1	1.15	0.39	0.95	0.89
2	1.66	0.44	1.00	1.05
3	-0.08	0.37	1.06	1.06
4	1.66	0.44	0.94	0.89
5	0.72	0.37	0.94	0.93
*6	0.86	0.38	1.11	1.25
7	1.15	0.39	0.97	0.98
*8	2.09	0.50	1.11	1.73
9	0.72	0.37	1.05	1.05
10	0.72	0.37	1.14	1.17
11	1.48	0.42	0.89	0.84
12	-0.81	0.41	1.05	1.08
13	0.05	0.36	1.02	1.04
14	0.58	0.37	1.01	1.04
*15	-0.50	0.38	1.20	1.23
16	0.72	0.37	1.02	1.10
17	-0.65	0.39	1.01	0.97
18	-0.81	0.41	1.07	0.99
*19	0.45	0.36	1.15	1.21
20	-2.66	0.74	1.06	1.12
*21	2.69	0.61	1.09	1.35
*22	2.69	0.61	1.11	1.49
23	0.86	0.38	1.09	1.18
24	0.72	0.37	0.96	0.95
25	0.86	0.38	1.05	1.01
26	-0.50	0.38	0.93	0.86
27	-1.37	0.46	0.89	0.81
28	-0.22	0.37	1.02	1.02
29	-0.08	0.37	0.98	0.98
30	1.48	0.42	0.99	1.02
31	0.05	0.36	0.91	0.92
*32	3.87	1.02	1.06	3.36
33	0.72	0.37	1.08	1.05
34	-0.98	0.42	1.06	1.12
35	-0.08	0.37	1.06	1.07
*36	-1.87	0.55	0.82	0.53
37	-0.08	0.37	1.04	1.04
*38	-0.98	0.42	1.00	1.21
*39	-1.87	0.55	0.87	0.62

*40	-1.60	0.50	0.88	0.69
41	-0.81	0.41	0.92	0.87
*42	-0.22	0.37	1.21	1.23
43	-0.08	0.37	0.94	0.95
44	0.72	0.37	0.98	0.92
45	0.86	0.38	0.97	0.90
46	0.19	0.36	0.92	0.90
47	-1.17	0.44	1.01	1.03
48	-1.17	0.44	1.06	1.11
49	-0.22	0.37	1.00	0.97
*50	-0.36	0.38	0.80	0.76
51	0.19	0.36	0.90	0.88
52	-1.60	0.50	0.95	0.93
*53	-1.17	0.44	0.84	0.74
54	0.58	0.37	1.05	1.08
55	-1.17	0.44	0.98	1.07
56	-0.81	0.41	0.90	0.83
57	-0.36	0.38	1.03	1.09
*58	-1.60	0.50	0.76	0.55
*59	-2.66	0.74	0.93	0.58
*60	-1.87	0.55	0.91	0.74

* *misfit items*

Table 5 displays the *fit* of the item data to the Rasch model. *Fit* refers to the degree of match between the pattern of observed responses and the modeled expectations.

It is also expressed in two ways: the pattern of responses observed for a person on each item is known as *person fit. ;* and the pattern of each item on all persons is known as *item fit* [22].

Using [23] as cited in [22] recommendations for a multiple-choice test-high stakes, the mean square range considered acceptable is 0.8 to 1.2. *Fit* statistics of less than 0.8 indicate overfit, which means the item has less variation than the model expects.

But in the table none of the infit mean square found to be lower than 0.8. However, there were items found to have outfit mean square higher than 1.2 which became the criteria for misfitting items. Seventeen (17) items were found to be misfit.

These are items 6, 8, 15, 19, 21, 22, 32, 36, 38, 39, 40, 42, 50, 53, 58, 59, and 60. Five of these items were also discarded using CTT approach and these are items 8, 21, 22, 32 and 42.

Among the six dimensions of CMAT, Language Structure, Verbal Analogy, Judgment and Comprehension, Mathematical Ability, Reading and Comprehension and Perceptual Acuity, each has items found to be misfit and these items are (6 &8), (15 & 19), (21 & 22) (32, 36, 38,&40), (42 &50) and (53, 58, 59 &60) respectively.

INPUT: 33 Persons 60 Items MEASURED: 33 Persons 60 Items 2 CATS

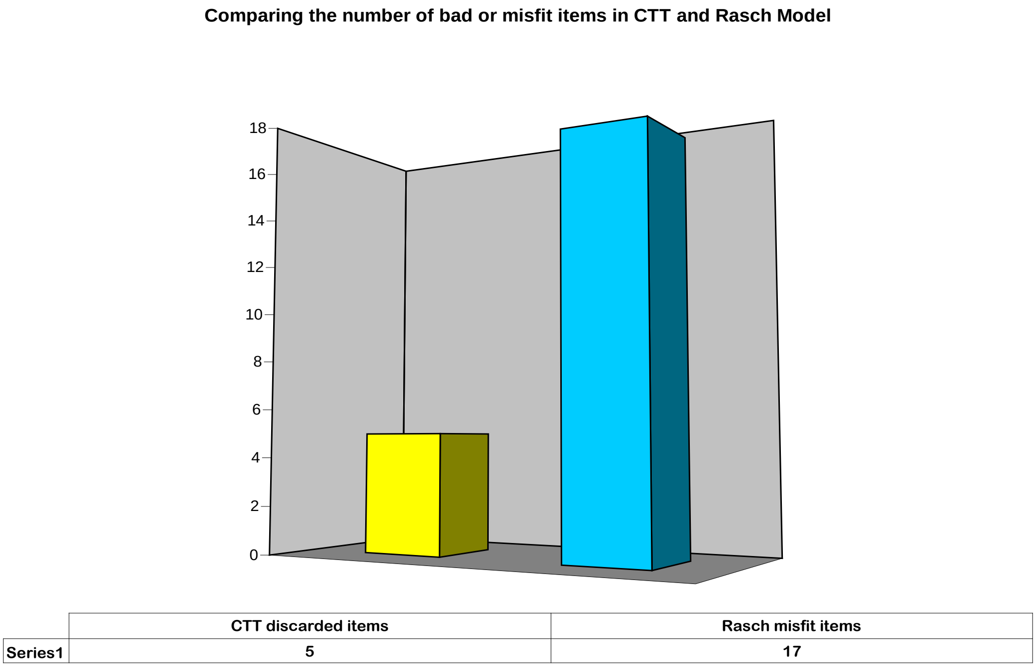
```
------------------------------------------------------------

     Persons - MAP - Items
<more>|<rare>
  +4      +
| 32
          |
          |
          |
          |
  +3      +
          |
          |
          |T 21    22
          |
          |
| 8
  +2      +
          |
          |
| 2     4
         T| 11    30
X  |
          |S
XXX  | 1     7
  +1      +
         S| 23    25    45    6
XXXX  | 10    16    24    33    44    5    9
XXX  | 14    54
XXX  | 19
XXXX  |
        X M|
X  | 46    51
   0    XXXX  +M 13    31
| 29    3    35    37    43
XXXX  | 28    42    49
         S| 50    57
X  | 15    26
| 17
X  | 12    18    41    56
XX  |
  -1      X T+  34    38
| 47    48    53    55
          |S
| 27
          |
| 40    52    58
          |
| 36    39    60
  -2      +
          |
          |
          |
          |
          |T 20    59
          |
          |
  -3      +
<less>|<frequ>
```

Figure 4: Person-Item Map

Figure 4 is the person-item map to show the location of the items and the person ability required to either pass or fail the item. From the left is the logit scale from -3 to +4. The logit scale is an interval scale in which all logit units have the same value. The highest values are the positive measures located on top of the map and the lowest values are located at the bottom. The more positive and higher the logit value, the more the difficulty of the item or the harder it is to endorse. On the other hand, the lower the logit value, the easier the item, the easier for it to endorse. Items that fit the model expectation are located in the middle vertical zone. Logits value between -2.0 to +2.0 are acceptable values in the map [22].

Figure 5 shows comparison of the number of discarded items both in CTT and the IRT (Rasch Analysis). There are 43 out of 60 items retained in Rasch Analysis because they are found to be *fit* or good items. Only 17 items were found to be misfit and these are meant to be deleted in the test. Unlike in CTT, in the Rasch analysis there is a middle way, which means the items can be revised or modified.

It is assumed in the Rash model the items that are found to fit have established its person ability reliability and item difficulty reliability indices. Hence, it is said that these fit items are good items and measures well.



	CTT discarded items	Rasch misfit items
Series1	5	17

Figure 5: The comparison of the number of items discarded in CTT and Rasch Analysis

Unidimensionality Coefficient

In the Rasch model, the context of unidimensionality refers to the meaningfulness of the estimates of person ability and item difficulty in the data matrix if each and every question contributes to the measure of a single attribute (Bond & Fox, 2001). The closer the value of the coefficient to 1.0, the more closely the data approximate unidimensionality. CMAT has unidimentionality index of 0.97. It means that as sets of measure variables in CMAT instrument form one attribute, it has only one underlying construct.

In the Rasch principal component analysis of model residuals conducted for the 60 item pool, it shows that 45.3% of the variance in the observation was accounted for by the Rasch dimension of item difficulty-person ability and the variance unexplained by first contrast is 5.6%.The chi-square goodness of fit is 2121.34 at p<.001.

4. Conclusion

The purpose of this study is twofold: 1. to identify good/fit and bad/unfit items to be retained or eliminated in the CMAT when both CTT and IRT Rasch model are usedg the process and whether the same items would be retained and eliminated whether one applies CTT or Rasch in developing and constructing tests items particularly multiple choice-dichotomous model; and 2. to identify the strength and limitations of CTT and IRT in test construction and development.

Both the CTT-based models and the IRT–Rasch model have strengths and limitations. The advantage of the CTT is the disadvantage of Rasch and vice versa. At this point, certain assumptions or criteria must be set when comparing the suitability of CTT and Rasch model in test development. CTT is advantageous when it comes to standardizing the entire test to a group of population, while the Rasch model is useful in standardizing two parameters - the person ability (the examinee's ability to answer the item correctly) and the *item difficulty* to establish the logits of the items. Generally, when student 1 and 2 belong to the same logit 2.28, and if the item is fit, they will have similar scores because they have similar capacity or probability that they will be successful in answering items with lower or negative value of logit. This refers to the replicability because Rasch model estimates the replicability of person placement across other items measuring the same construct. The estimate is based on the same concept as Cronbach's alpha [22].

t is meaningful to compare these two models in test development because they have different perspectives of treating the tests as a whole. At a certain point, it is exhausting to compare CTT to Rasch model but to compare the two, there must be a point wherein they are similar in context: 1. the length of the test, 2. the number of samples, 3. the procedure of item analysis, 4 the process of treating item difficulty 5.the person ability to answer the item right 6. the process of retaining the good/ fit items and eliminating the bad or misfits items, are entirely different approach. CTT and Rasch model have opposite baseline principles in test construction. Although these approaches are entirely different, the final items would still remain the end-product of the instrument. Hence, in this study, the best way that CTT and Rasch may be analyzed is in the final items retained in the instrument granting the SEM is closer to zero. Looking into the CMAT items, apparently, the discarded items in CTT are also found to be misfit items in Rasch model, this is helping the test developers to use Rasch model to refine their instrument.

As shown in the results, there are more misfit items deleted using Rasch model as compared to CTT, and the former seemed to be stricter in that sense as compared to the latter. Since the CTT and the Rasch model- have inherently recurring issues and differences, the best option that test developers can carry out is to choose between the two models depending on the purpose and the function of the test. If the need is to establish norm-referenced criterion and determine the reliability and validity, CTT may be employed. However, if the test developer prefers to establish the reliability indices of the *item difficulty* based on the *person ability* to answer the test correctly using probability measure, regardless of the target population, Rasch analysis should be employed. On the other hand, if the purpose of the instrument is to generate item banking for long tests, CTT is more suitable. However, if, as a test developer thinks that a short test is enough to measure intelligence, aptitude, achievement and other cognitive tests, Rasch model will be most appropriate. If the sample is over 1000, (measuring the entire population of all high school in the Philippines) CTT is still a better choice. However, if the test developer prefers a shorter but reliable instrument in just a small population (class size), Rasch model is applicable.

Both the CTT and the Rasch Model are good grounds for deciding what items to retain and discard, using both CTT and Rasch. For example in a 100-item of test, using CTT,10 items are discarded, and 15 in Rasch model, , in order to satisfy both the reliability in both conditions, simply sum up 10 and 15, (if the items are different) and include only those items that are good/ fit in the final draft of the instrument. But in the result of CMAT, the

5 items deleted in CTT are also found bad items in the 17 misfit items using Rasch model. Apparently, there are still 43 good or fit items left out of 60 items in CMAT. Generally, it happens in several cases, then, it is preferred to use Rasch model from the start. However, if there are different set of items are deleted in CTT and another set of misfit items to be deleted in the Rasch model, simply sum up these bad/ misfit items to satisfy both CTT and the the Rasch Model conditions in order to generate good set of items for item pooling.

One further note towards test development and test construction and validation, the Rasch model is only one step towards the goal of the creation of reliable or more stable and valid educational and psychological measures whether the data is dichotomous (data answerable by yes or no) or polytomous (data answerable in scale such as Likert). However, Rasch model alone is not the solution to all the instrument and measurement problems, it is just a mathematical model. The test development and construction is a composition of psychometric properties; good constructed items, validity and reliability and well defined constructs that will measure latent or unobserved trait or ability. The bottom line of the discussion still relies on how the well the items are constructed, developed and validated.

Acknowledgements

The author would like to acknowledge the support and assistance of Dr. Alexa P. Abrenica, Dr. Carlo Magno for the valuable comments and suggestions in writing the manuscript, Director Marife M. Mamauag and Mr. Johnny Amora and and Mr. Sonny Magno for helping her carry out the data into WINSTEPS program and Cavite State University Indang Campus for being supportive.

References

[1] C. Cantell,. Item response theory: Understanding the one-parameter Rasch Model. A paper presented at the annual meeting of the Southwest Educational Research Association Austin TX, 1997.

[2] R.J. Cohen. Exercises in psychological testing and assessment: An Introduction to tests and measurements 6th ed., New York: McGraw-Hill International. 2005.

[3] R.J. Cohen, & M.E. Swerdlik. Psychological testing and assessment: An Introduction to tests and measurements 6th ed., New York: McGraw-Hill International. 2007.

[4] S.E. Embretson & S.P. Reise. Item response theory for psychologist. London: Lawrence Erlbaum Associates, Publishers, 2000.

[5] F. Xitao. Item response theory and classical test theory: an empirical comparison of their item/person statistics. Educational and Psychological Measurement Sage Publication, vol. 58 n3 pp. 357 (25) June 1998.

[6] J.E. Gustafsson. An introduction to Rasch measurement model. Paper presented at the Nordic Researcher's Course "Rasch Models' in the social and behavioral sciences." Princeton, NJ, 1990.

[7] Henson, R.. Understanding the one-parameter Rash Model of item response model of the item response theory. A paper presented at the annual meeting of the Southwest Educational Research Association San Antonio,TX. 1999

[8] Kubiszyn, T. & Borick, (2004) Educational testing and measurement Classroom Application and Practice 7thed. New York: John Wiley& Sons, Inc. ,2004.

[9] Linacre J.M. True-score reliability or Rasch statistical validity? Rasch Measurement Transaction 9:4 p. 455-6, 1996.

[10] Lord, F. M.. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum. 1980

[11] Magno, C. & Hai, CY. The Application of a One- parameter IRT Model on A Test of Mathematical Problem Solving.

[12] Reeve, B. B. Item response theory modeling in health outcomes measurement. Expert Review of Pharmacoeconomics and Outcomes Research, 3 (2), 131-145. 2003

[13] B. B. Reeve, & P. Fayers, P. .Applying item response theory modeling for evaluating questionnaire item and scale properties. In P. Fayers and R. D. Hays (Eds.), Assessing Quality of Life in Clinical Trials: Methods of Practice. 2ndEdition. Oxford University Press. 2005. pp.55-73.

[14] B.B. Reeve, R.D. Hays, J.B. Bjorner, K.F. Cook, P.K. Crane, J.A. Teresi, D. Thissen, D. A.Revicki, D.J. Weiss, R.K. Hambleton, H. Liu, R. Gershon, S.P. Reise, J.S. Lai, D. Cella,. Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Medical Care.45(5) S22-S31. 2007.

[15] Wiberg, M. Classical Test theory vs. Item Response theory. EM No.50, 2004.

[16] Wright, B. D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems ,Princeton, N.J.: Educational Testing Service.1967.

[17] Wright, B. D. & N. Panchepakesan,. A procedure for sample free item analysis. Educational and Psychological Measurement, 1969. 29, 23-37.

[18] B. D. Wright & R. J. Mead. Calfit: Sample-free item calibration with a Rasch measurement model. Research Memorandum, No. 18. Statistical Laboratory, Department of Education, University of Chicago, 1975

[19] B. D. Wright & G. A. Douglas. Better procedures for sample-free item analysis. Research Memorandum, No. 20 .Statistical Laboratory, Department of Education, University of Chicago. 1975

[20] B. D. Wright & M.H. Stone. Best Test Designs and Self – Tailored Testing. Research Memorandum , No. 19. Statistical Laboratory, Department of Education, University of Chicago., 1975

[21] B. D. Wright & M.H. Stone. Best Test Designs. Chicago: Mesa Press, 1979.

[22] T.G. Bond, & C.M. Fox. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Rasch Measurement Transactions, 2001, 15:1 p.790.

[23] B.D.Wright & JM Linacre. Reasonable mean-square fit values. Rasch Measurement Transactions, 1994, 8:3 p.370.