ISSN 2348-3156 (Print)
International Journal of Social Science and Humanities Research ISSN 2348-3164 (online)
Vol. 3, Issue 2, pp: (22-28), Month: April - June 2015, Available at: www.researchpublish.com

Construct Validity of Mathematics Test Items Using the Rasch Model

ALIYU, R.TAIWO

Department of Guidance and Counselling (Measurement and Evaluation Units)
Faculty of Education, Delta State University, Abraka, Nigeria

Abstract: **The paper addresses the procedures to adopt in validating Mathematics Test Items (MTI). The content validity was examined based on some experts' judgment on the development of the items. The analysis of the 100 self-developed MTI with a sample of 200 testees was based on winsteps analysis. The result showed that the 86 items not only met the Rasch model assumption of measurement construct but also demonstrated good psychometric properties.**

Keywords: **difficulty, Rasch model, unidimensionality and construct validity.**

1. INTRODUCTION

One of the ultimate purposes in educational measurement is to estimate testee's ability in a particular subject. This measurement always involves assigning numerical numbers to certain traits or characteristics using a tool (Aliyu & Ocheli, 2012). For physical traits, such as height, the process of assigning numbers can be done directly using a ruler. However, psychological traits such as ability or proficiency are constructs. They are unobservable but can be measured indirectly using a tool called test. The design of tests to measure constructs, however, presents several problems. Since the measurement of psychological constructs is always done indirectly, there is always the possibility that researchers will select different types of behavior to measure the same construct. As a consequence, different inferences will be concluded. Lack of well-defined units in the measurement scale also poses problem. For example, an examinee who is unable to answer any test item does not mean that he or she has "zero" ability. Instead, all the items have difficulty index which is more than the examinee's ability. The study of measurement problems and methods to overcome them is known as test theory. Test theories relate observable traits (such as *test score*) with unobservable traits (such as *ability* or *proficiency*) for a measured construct using mathematics model (Odili, 2010).

The first established test theory is called the Classical Test Theory (CTT). The CTT revolves around concepts of true score, measurement error and index of test reliability. CTT relates observable trait (the test score, X) with the unobservable trait (the person's true ability on the characteristics, T) with the following equation: X = T + E, where E = measurement error (Osadebe, 2010). Item Response Theory (IRT), meanwhile, relates responses to test items (observable trait) to unobservable traits through models that specify how both trait level and item properties are related to person's item response (Embretson, 2000). Three IRT models have been developed. They are named for the number of parameters they use to estimate examinee ability.

One parameter model, also known as the Rasch Model, uses only single parameter, namely item difficulty to estimate an unobservable trait of a particular examinee. The two-parameter and three-parameter models are also widely used, especially in large scale assessment (Downing, 2003). The two-parameter adds an item discrimination parameter to the item difficulty, whereas the three parameter model adds a 'guessing' parameter to item difficulty and item discrimination. Hambleton and Swaminathan (1985) provide substantial description of the two-parameter and three-parameter models as well as item response theory as a whole.

One of the major limitations of the CTT is that the item statistics (the difficulty index, *p*-value) and (the discrimination index, *r*-values) which are very essential in the application of CTT are sampled dependent. These limitations are

addressed and overcome in IRT. When its assumptions have been satisfied, IRT provides (1) examinee ability measures that are independent on the particular sample of test items chosen, (2) item statistics that are independent of sample of examinee drawn, and (3) fit statistics indicating the precision of the estimated ability for each examinee and precision of each item. A classic article by Wright (2002) provides readers with detailed explanation to invariant person and item parameter known as '*examinee-free*' test calibration and '*item-free*' examinee measurement. With ability estimates being invariant; IRT provides a way of comparing examinee even though they take a different test. All IRT models offer invariant properties for estimation of item and examinee parameters.

According to Ahmad (2012), the choice of appropriate model depends on the type of test questions and their scoring. Another important consideration is that, in practice, the choice of models depends on the amount of data available. The larger the number of parameter is, the more data are needed for parameter estimation, thus requiring more complex calculation and interpretation. In this case, Rasch Model has some special properties that make it attractive to users. Rasch Model involves fewest parameters; therefore, it is easier to work with (Downing, 2003). Wright (1990) gives more influential explanation in favor of Rasch Model compared to a three-parameter model. These two models are opposite in philosophy and in practice. The three-parameter model will adjust to adapt whatever type of data (includes invalid responses). The Rasch model however has tight standards in controlling the data. Unlike the three-parameter model, invalid responses such as guessing on item will not be accepted. It is described as unreliable person reliability. Critics of the Rasch Model often regard the model as having strong assumptions that are difficult to meet. However, these are values that make Rasch Model more appropriate in practice.

One major problem in measurement lies in the interaction between the person being measured and the instrument involved. Performance of a person is known to be dependent on which instrument is used to measure his or her trait. However, this shortcoming is circumvented by procedure of conjoint measurement in Rasch Model. Olaleye & Aliyu (2013) explain that in conjoint measurement, the unit of measurement is not the examinee or the item, but rather the performance of an examinee relative to a particular item. If $ß_n$ is an index for ability for examinee *n* on the trait being measured, and if δ_i is an index for the difficulty of the item *i* which relates to the trait being measured, then the unit of measurement is neither $ß_n$ nor δ_i but rather $(ß_n - \delta_i)$, which is the difference between the ability of the examinee and the difficulty of the item. If the ability exceeds the item difficulty, then it is expected that the examinee will answer the item correctly. In contrast, if the difficulty exceeds the ability, then it is expected that the examinee will answer incorrectly. In education, response on a particular item is always in uncertainties. Therefore, probabilistic approach has to be employed when explaining what happens when an examinee takes an item. Probabilities of correct response are between 0 and 1 and it does not permit proportion of correct answer to be expressed in interval scale. To overcome these constraints, logistic transformation, which involves taking the natural logarithm, is used. As a final product, it can be shown that the probability of person *n* has correct response to item I is given by (Ahmad, 2012)

$$\Pr\{X_{ni} = 1\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}},$$

Rasch Model offers procedure to transform test score into interval-scale measure (score) in log-odd or *logits* unit. Earlier works clearly showed a need for interval-scaled measures in measurement of an intended construct (Andrich, 1999). Application of summated score (such as number of correct answers) have been strongly opposed due to the fact that it is highly unlikely that examinee score can be interpreted accurately, cannot determine how one's score is different from other examinee and that the difference between two scores is not reliable when different scoring scheme is in used (Bond, 2001). In Rasch Model analysis, two important parameters usually discussed are item difficulty and examinee ability. Item difficulty measure is an estimate of an item's underlying difficulty calculated from the number of examinee who succeeds in that item. Examinee's ability measure, on the other hand, is an estimate of his or her underlying ability based on performance on a set of items. In order for the Rasch Model measurement to have the '*examinee-free*' item difficulty and '*item-free*' examinee ability measurement, two important assumptions must be met. Firstly, the data must meet the unidimensionality assumption, that is, they represent a single construct (Bond, 2001). Secondly, Rasch Model requires that the data must fit the model (Bond, 2001). In addition, it is also imperative to provide evidence on the psychometric properties of the test used from the framework of Rasch Model analysis. Based on the foundation laid by Messick, (1993) two major threats to construct validity that are under investigation are construct-irrelevant variance and construct under-representation. The former relates to the irrelevant variances that contaminate measurement of the main construct while in

the latter, the measurement fails to include important sub-dimensions of the construct. In short, construct validity requires nothing irrelevant be added while at the same time nothing important should be left out in assessing a construct. Within the framework of Rash Measurement Model, Baghari (2008) suggests that construct-irrelevant variance can be assessed by examining both dimensionality and fit of the measurement while significant gaps between the subsequent items provide indication of construct under-representation.

2. STATEMENT OF PROBLEM

The poor achievement of students in Mathematics has been of great concern in the society at large. Aliyu & Ocheli (2012) have responded to the declining trend in student's performance by advocating for the new approach of analyzing test data. The commonest method of assessment of students' performance has always been the classical test theory (CTT) which lack objectivity in the cognitive and psychomotor traits of the candidates. Therefore, the statement of problem if put in a question form is: How suitable is the development and validation of a MTI able to determine students' achievement in Mathematics using the Rasch model?

Research Questions:

This study therefore, attempts to answer the following questions.

1. What are the infit and outfit indices of MTI item using the Rasch model?
2. What are the validity and the reliability estimates of MTI items using the Rasch model?

Purpose of the study:

The main purpose of the study is to develop and validate Mathematics Test Item using the Rasch model. In light of the preceding discussion, the present study is aimed to (1) examine the extent to which a set of test to measure Mathematics achievement meets Rasch Model expectation, and (2) provide evidence of adequate psychometric properties of the Mathematics Test Item (MTI).

Significance of the study:

The study intends to make lecturers see that the test that measure achievement in Mathematics is needed. This test scores will directly convey level of competence in defined Mathematics domain. The study equally intends to be a guild to test developers in the development and validation process.

3. RESEARCH METHOD

This study was designed to be an instrumentation research, which is non-experimental since it involved the development and validation of an instrument called MTI. This was used in measuring students' achievement.

Population:

The population of the study was made up of all year 2 students in Delta State University.

Sample and Sampling Techniques:

A total of 200 students were used for the final administration of the instrument. A multi-stage sampling technique was adopted. Two out of the four Campuses of the University were randomly selected. Simple randomly sampling was equally used for selecting the four (4) departments, 2 from each in the two (2) campuses while non-proportionate stratified random sampling were adopted in selecting the students in the selected schools(50) each to arrive at the needed sample of 200 for the study.

Development and validation of instrument:

The instrument for this study was a self-developed Mathematics Test Item (MTI). It is a 150 items drawn from the content of the University students' Mathematics curriculum as structured into their modules. A table of specification of only 3 cognitive levels, knowledge, comprehension, and higher as advised by other researchers (Anigbo, 2012, Adedoyin et al 2008, Opasina 2009) were considered. In this arrangement, the 4 higher level objectives (application, analysis, synthesis and evaluation) were grouped together. Experts in the field of Mathematics verified the instrument. These were done to ensure both content and face validity of the instrument. Some items were deleted while some were reconstructed which led to the emergence of 100 items from the vetting exercise and were trial tested. They were administered to 50 students (30 boys and 20 girls) who were not part of the sample used.

ISSN 2348-3156 (Print)
International Journal of Social Science and Humanities Research ISSN 2348-3164 (online)
Vol. 3, Issue 2, pp: (22-28), Month: April - June 2015, Available at: www.researchpublish.com

Reliability of the instrument:

A KR-20 reliability method was employed in testing the reliability coefficient of the instrument. The value obtained was 0.75. On the basis of the calculated reliability coefficient, the instrument was considered reliable for the study and administered to the 200 samples.

Analysis:

In this study, a Rasch Model software, WINSTEPS version 3.75 (Aliyu, 2013) is used. In WINSTEPS, the measures are determined through iterative calibration of both person and item using the Mathematics Achievement test. In WINSTEPS, the infit mean square (MNSQ) and outfit MNSQ provide indications of the discrepancies between the data and model's expectations. This study adopts the range of acceptable fit between 0.7 – 1.3 for both fit indices as suggested by (Bonds & Fax, 2001). Psychometric properties of the test were tested in terms of reliability and validity of the measures, meaning and interpretation were given. Rasch analysis provides reliability indices for both item and examinee's measure. High reliability for both indices are desirable since they indicate a good replication if the comparable items/examinees are employed.

4. ANALYSIS AND RESULTS

Item STATISTICS: CORRELATION ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
45	110	200	.28	.14	1.10	4.3	1.11	4.4	-.24	.13	48.0	56.6	I0045
7	51	200	1.58	.16	1.05	.6	1.08	.9	-.07	.12	74.5	74.5	I0007
37	114	200	.20	.14	1.06	2.0	1.06	1.9	-.06	.13	50.0	57.8	I0037
32	143	200	-.45	.16	1.05	.7	1.07	.9	-.06	.12	71.5	71.5	I0032
27	64	200	1.25	.15	1.05	.9	1.06	1.0	-.05	.13	68.0	68.0	I0027
23	103	200	.42	.14	1.05	2.6	1.05	2.5	-.05	.14	48.5	55.5	I0023
96	155	200	-.77	.17	1.04	.4	1.07	.7	-.05	.11	77.5	77.5	I0096
17	78	200	.94	.15	1.05	1.3	1.05	1.4	-.04	.13	61.0	61.5	I0017
25	97	200	.55	.14	1.05	2.3	1.05	2.3	-.03	.14	50.0	55.6	I0025
29	170	200	-1.27	.20	1.02	.2	1.07	.5	-.02	.09	85.0	85.0	I0029
49	81	200	.88	.15	1.04	1.2	1.04	1.3	-.01	.13	57.5	60.3	I0049
50	96	200	.57	.14	1.04	2.0	1.04	2.0	-.01	.14	49.5	55.8	I0050
16	98	200	.53	.14	1.04	1.8	1.04	1.8	.01	.14	51.5	55.5	I0016
26	65	200	1.23	.15	1.03	.6	1.04	.7	.01	.13	67.0	67.5	I0026
15	107	200	.34	.14	1.03	1.6	1.03	1.5	.01	.14	50.0	55.9	I0015
99	147	200	-.55	.16	1.02	.3	1.04	.5	.02	.12	73.5	73.5	I0099
2	132	200	-.19	.15	1.03	.6	1.03	.6	.03	.13	66.5	66.0	I0002
6	71	200	1.09	.15	1.03	.6	1.03	.7	.03	.13	65.5	64.6	I0006
31	118	200	.11	.15	1.03	.9	1.03	.9	.03	.13	57.5	59.3	I0031
28	72	200	1.07	.15	1.02	.6	1.03	.6	.04	.13	65.0	64.2	I0028
92	148	200	-.58	.16	1.02	.3	1.02	.3	.04	.12	74.0	74.0	I0092
36	83	200	.83	.14	1.02	.8	1.02	.7	.05	.14	51.5	59.5	I0036
78	158	200	-.86	.17	1.01	.2	1.02	.2	.05	.11	79.0	79.0	I0078
24	70	200	1.12	.15	1.02	.4	1.02	.5	.06	.13	64.0	65.1	I0024
89	155	200	-.77	.17	1.01	.2	1.03	.3	.06	.11	77.5	77.5	I0089
52	176	200	-1.53	.22	1.00	.1	1.02	.2	.06	.09	88.0	88.0	I0052
93	150	200	-.63	.16	1.01	.1	1.03	.4	.07	.12	75.0	75.0	I0093
86	144	200	-.48	.16	1.01	.2	1.02	.3	.07	.12	72.0	72.0	I0086
43	96	200	.57	.14	1.02	.9	1.02	1.0	.07	.14	55.5	55.8	I0043
82	148	199	-.60	.16	1.01	.2	1.02	.3	.07	.12	74.4	74.4	I0082
48	128	200	-.10	.15	1.01	.4	1.02	.4	.07	.13	64.0	64.0	I0048
100	154	200	-.74	.17	1.01	.1	1.02	.3	.08	.11	77.0	77.0	I0100
61	155	200	-.77	.17	1.01	.1	1.01	.1	.09	.11	77.5	77.5	I0061
59	144	200	-.48	.16	1.01	.2	1.00	.1	.09	.12	72.0	72.0	I0059
55	166	200	-1.12	.19	1.00	.1	1.00	.0	.09	.10	83.0	83.0	I0055
94	142	200	-.43	.16	1.01	.2	1.00	.1	.09	.12	71.0	71.0	I0094
60	146	199	-.55	.16	1.01	.1	1.01	.1	.09	.12	73.4	73.4	I0060
35	79	200	.92	.15	1.01	.3	1.01	.3	.10	.13	59.5	61.1	I0035
97	145	199	-.52	.16	1.00	.1	1.01	.2	.10	.12	72.9	72.9	I0097
22	72	200	1.07	.15	1.01	.2	1.01	.2	.10	.13	64.0	64.2	I0022
98	142	200	-.43	.16	1.00	.1	1.00	.1	.11	.12	71.0	71.0	I0098
91	148	200	-.58	.16	1.00	.0	1.01	.1	.11	.12	74.0	74.0	I0091
90	148	200	-.58	.16	1.00	.0	1.00	.1	.11	.12	74.0	74.0	I0090
88	137	199	-.32	.15	1.00	.0	1.01	.1	.11	.12	68.8	68.8	I0088
51	178	200	-1.63	.23	.99	.0	.97	-.1	.12	.08	89.0	89.0	I0051
87	129	200	-.12	.15	1.00	.0	1.01	.1	.12	.13	64.5	64.5	I0087
53	174	200	-1.44	.21	.99	.0	.99	.0	.12	.09	87.0	87.0	I0053
73	146	199	-.55	.16	1.00	.0	1.00	.0	.12	.12	73.4	73.4	I0073
85	140	200	-.38	.16	1.00	.0	1.00	.0	.13	.12	70.0	70.0	I0085
95	131	200	-.17	.15	1.00	.0	1.00	.0	.13	.13	66.5	65.5	I0095
69	160	200	-.92	.18	.99	.0	.99	-.1	.13	.11	80.0	80.0	I0069
8	74	200	1.03	.15	1.00	.0	1.00	.0	.13	.13	62.5	63.2	I0008
21	96	200	.57	.14	1.00	.0	1.00	.0	.14	.14	55.5	55.8	I0021
81	139	200	-.35	.15	1.00	.0	.99	-.1	.14	.12	69.5	69.5	I0081
58	151	199	-.68	.17	.99	.0	.99	-.1	.14	.11	75.9	75.9	I0058
10	102	200	.44	.14	1.00	-.2	1.00	-.2	.15	.14	60.0	55.5	I0010
39	106	200	.36	.14	.99	-.3	.99	-.2	.16	.14	61.5	55.7	I0039
5	87	200	.75	.14	.99	-.3	.99	-.3	.16	.14	57.5	58.0	I0005

ISSN 2348-3156 (Print)
International Journal of Social Science and Humanities Research ISSN 2348-3164 (online)
Vol. 3, Issue 2, pp: (22-28), Month: April - June 2015, Available at: www.researchpublish.com

66	150	200	-.63	.16	.99	-.1	.97	-.3	.16	.12	75.0	75.0	I0066
14	88	200	.73	.14	.99	-.3	.99	-.4	.17	.14	56.5	57.7	I0014
30	91	200	.67	.14	.99	-.4	.99	-.4	.17	.14	60.0	56.8	I0030
57	152	200	-.69	.17	.99	-.1	.97	-.3	.17	.11	76.0	76.0	I0057
77	148	200	-.58	.16	.99	-.1	.97	-.3	.17	.12	74.0	74.0	I0077
80	138	200	-.33	.15	.99	-.2	.98	-.3	.17	.12	69.0	69.0	I0080
84	144	200	-.48	.16	.99	-.2	.98	-.3	.17	.12	72.0	72.0	I0084
12	92	200	.65	.14	.99	-.5	.99	-.5	.18	.14	60.5	56.6	I0012
18	93	200	.63	.14	.99	-.5	.99	-.5	.18	.14	61.0	56.4	I0018
63	147	200	-.55	.16	.99	-.2	.97	-.3	.18	.12	73.5	73.5	I0063
42	52	200	1.55	.16	.98	-.2	.98	-.2	.18	.12	74.0	74.0	I0042
13	102	200	.44	.14	.99	-.6	.99	-.6	.18	.14	59.0	55.5	I0013
65	157	199	-.86	.17	.98	-.1	.97	-.3	.18	.11	78.9	78.9	I0065
1	95	200	.59	.14	.99	-.6	.99	-.7	.18	.14	55.0	55.9	I0001
74	147	200	-.55	.16	.98	-.2	.97	-.4	.19	.12	73.5	73.5	I0074
68	141	199	-.42	.16	.98	-.2	.97	-.4	.19	.12	70.9	70.8	I0068
19	73	200	1.05	.15	.98	-.4	.99	-.3	.19	.13	65.0	63.7	I0019
54	170	200	-1.27	.20	.98	-.1	.96	-.3	.19	.09	85.0	85.0	I0054
79	132	200	-.19	.15	.98	-.3	.97	-.5	.20	.13	66.5	66.0	I0079
62	140	200	-.38	.16	.98	-.3	.97	-.4	.20	.12	70.0	70.0	I0062
9	85	200	.79	.14	.98	-.6	.98	-.7	.20	.14	59.5	58.7	I0009
71	158	200	-.86	.17	.98	-.2	.96	-.4	.20	.11	79.0	79.0	I0071
75	148	200	-.58	.16	.98	-.3	.96	-.4	.21	.12	74.0	74.0	I0075
83	125	200	-.03	.15	.98	-.5	.97	-.6	.21	.13	62.5	62.6	I0083
47	100	200	.49	.14	.98	-1.1	.98	-1.1	.21	.14	61.0	55.5	I0047
3	79	200	.92	.15	.98	-.6	.98	-.6	.21	.13	64.5	61.1	I0003
44	99	200	.51	.14	.98	-1.2	.98	-1.2	.22	.14	59.0	55.5	I0044
56	155	199	-.80	.17	.97	-.2	.95	-.4	.22	.11	77.9	77.9	I0056
20	127	200	-.08	.15	.97	-.6	.97	-.7	.23	.13	63.5	63.6	I0020
67	158	200	-.86	.17	.97	-.2	.94	-.5	.23	.11	79.0	79.0	I0067
72	148	200	-.58	.16	.97	-.3	.95	-.6	.23	.12	74.0	74.0	I0072
4	98	200	.53	.14	.97	-1.5	.97	-1.5	.24	.14	58.5	55.5	I0004
70	146	200	-.53	.16	.97	-.4	.95	-.6	.24	.12	73.0	73.0	I0070
34	72	200	1.07	.15	.97	-.7	.96	-.7	.25	.13	65.0	64.2	I0034
64	143	200	-.45	.16	.97	-.5	.95	-.7	.25	.12	71.5	71.5	I0064
40	64	200	1.25	.15	.96	-.6	.96	-.6	.26	.13	68.0	68.0	I0040
38	82	200	.86	.15	.97	-1.1	.96	-1.1	.26	.14	61.0	59.9	I0038
11	124	200	-.01	.15	.96	-1.0	.96	-1.0	.27	.13	63.0	62.1	I0011
76	140	199	-.40	.16	.96	-.6	.94	-.9	.28	.12	70.4	70.3	I0076
46	88	200	.73	.14	.95	-1.9	.95	-1.9	.31	.14	61.5	57.7	I0046
41	88	200	.73	.14	.94	-2.5	.94	-2.5	.36	.14	65.5	57.7	I0041
33	86	200	.77	.14	.93	-2.7	.92	-2.7	.40	.14	65.0	58.4	I0033
MEAN	120.8	199.9	.00	.16	1.00	.0	1.00	.0			67.7	67.7	
S.D.	33.2	.3	.76	.02	.03	.9	.03	.9			9.3	8.9	

In answering the **RQ 1**, the **infit** and **outfit** columns for both **MNSQ** and **ZSTD** show the indices. The table equally shows that item **7** is the most difficulty item in the test. The difficulty of this item is estimated to be **1.58logits** with the standard error of **0.16** while item **51** is the easiest with **-1.63logits** and standard error of **0.23**. Table 1 equal indicates that items **45, 7, 37, 32, 27, 23, 96, 17, 25, 29, 49, 50, 41** and **33** should be omitted, deleted or revised because of lack of fit to the model. These items are measuring something other than the intended content and construct. They are construct irrelevant. There are 86 items that met the Rasch model assumption which is an indication of **unidimensionality** of the **MTI**.

SUMMARY OF 200 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	60.4	99.9	.49	.22	1.00	.0	1.00	.0
S.D.	5.7	.7	.28	.01	.10	1.3	.13	1.3
MAX.	75.0	100.0	1.24	.24	1.30	3.4	1.38	3.3
MIN.	45.0	90.0	-.23	.21	.70	-4.0	.64	-3.9

REAL RMSE	.22	TRUE SD	.16	SEPARATION	.73	Person RELIABILITY .55
MODEL RMSE	.22	TRUE SD	.17	SEPARATION	.77	Person RELIABILITY .57
S.E. OF Person MEAN = .02						

Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .55

In answering **RQ 2**, the summary statistics table is considered. The separation index of the persons is **0.73** which translates to a person strata index of **3.4**. The strata index shows the number of distinct ability levels which can be identified by the test (Reza & Baghaei, 2011). The minimum person strata index is 2 which means that the test is able to distinguished between at least 2 strata of persons namely, high-ability and low-ability persons. A reliability index of at least **0.50** is required for a separation index of 1. The moderate reliability, separation and strata indices for this test are as a result of the low standard deviation of the person abilities. The Crobach Alpha (KR-20) person raw score test reliability

ISSN 2348-3156 (Print)
International Journal of Social Science and Humanities Research ISSN 2348-3164 (online)
Vol. 3, Issue 2, pp: (22-28), Month: April - June 2015, Available at: www.researchpublish.com

of **0.55** was moderate, indicating that it was likely the ordering of the examinees ability can be replicated since most of the variance was attributed to true variance of the Mathematics Test Items (MTI).

SUMMARY OF 100 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	120.8	199.9	.00	.16	1.00	.0	1.00	.0
S.D.	33.2	.3	.76	.02	.03	.9	.03	.9
MAX.	178.0	200.0	1.58	.23	1.10	4.3	1.11	4.4
MIN.	51.0	199.0	-1.63	.14	.93	-2.7	.92	-2.7

REAL RMSE	.16	TRUE SD	.74	SEPARATION	4.68	Item	RELIABILITY	.96
MODEL RMSE	.16	TRUE SD	.74	SEPARATION	4.71	Item	RELIABILITY	.96
S.E. OF Item MEAN = .08								

One could investigate the representativeness of the test items too by checking the separation index. The separation shows the spread of the items along the variable void of gaps and targeted to person ability (Wright, 1988). The minimum index for item separation and item strata is 2. Therefore, the separation value for this test is **4.68**. The item reliability **0.96** is a very good one which shows that the items are very reliable for administration. There is a very wide spread of difficulty in the items as the standard deviation of item difficulty estimates is **0.76logits** and the separation is **4.68**.Thus, one can rely on the representativeness of the test items.

5. FINDINGS AND DISCUSSION

The Rasch analysis as presented in Table 1 found both means of infit MNSQ and outfit MNSQ were close to the expected value of **1.00**. Inspection with individual items showed that infit MNSQ values ranged from **0.90** to **1.11** while outfit MNSQ values ranged from **-1.90** to **1.20**. The results supported the following: (1) the unidimensionality assumption of the construct validity was met, and (2) the scores demonstrated little variation from model expectation – that there was evidence of consistency between 200 examinees' response and 100 items on the scale and the model's expectations. Reliability of item difficulty measures were high **(.96)** suggesting that the ordering of item difficulty was replicable with other comparable sample of examinee. From the findings, threat regarding construct irrelevant-variance was minimum based on the dimensionality test as well as the within-range fit indices. In all, there are **86** items that fit the Rasch model with an indication of **unidimensionality**.

6. CONCLUSION AND RECOMMENDATIONS

The present study extends the understanding of how Rasch Model framework could be used in test development to measure certain construct. Using the previously established processes of numerous scholars that link Rasch-based analyses to the six facets of Messick's (1989) construct validity, the current study used the results of these analyses to provide validity arguments in an evaluation of Mathematics Test Items (MTI). The test exhibited few negative point-measure correlations and has very few misfitting items. The test did exhibit a fairly low mean score although test-takers' abilities were nonetheless, reasonably well spread across items. The limitation of the current study, while attempting to provide validity evidence, did not include such analyses as, differential item or test functioning, unexpected response or item distractor analyses or person-item-map.

These should most certainly be explored in more detail to determine if there are any items that are causing unexpected response patterns either across groups or across sections of the test. Baghaei and Amrahi (2011) noted that it is not entirely reasonable to simply sum different parts of a test if each part is measuring a different dimension. The use of Rasch Model offers opportunity to deal with core measurement issues such as construct validity as well as providing richer interpretation regarding examinee performance. Theoretically, this study has added more evidence in favor of the Rasch Model as having the capacity to resolve some of the rudimentary issues in measurement. However, in order for construct validity to hold, the model requires more evidence. Test developers would have to have a thorough understanding of the measured construct especially information on relative difficulties of the items so that they can conceptualize the measured construct. This Rasch analysis has provided useful information which not only can be used for future developments, modification and monitoring achievement assessments, but also for establishing a process of validating pedagogical assessment.

REFERENCES

[1] Adedoyin,O.O, Nenty, H. J & Chilisa, B. (2008). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. Educational Research and Review vol. 3(2) 83-93

[2] Ahmad, Z.K. & Nordin, A. (2012). Advance in Educational Measurement: A Rasch Model Analysis of Mathematics Proficiency Test. International Journal of Social Science and Humanity, vol. 2, No. 3

[3] Aliyu, R. T. (2013). Development and validation of Mathematics Achievement Test using the Rasch Mode. An unpresented Ph.d thesis in Delta state university, Abraka.

[4] Aliyu, R.T. & Ocheli, E.C. (2012). Development and validation of College Mathematics test (CMAT) with Item Response Theory (IRT) models in attaining quality education for national value. A paper presented in the international conference held at Delta state University (DELSU) between 13-17, November, 2012.

[5] Andrich, D. (1992). "The application of an unfolding model of the PIRT type to measurement of attitude". Applied Psychological Measurement, vol. 12, pp. 33-5.

[6] Anigbo, L.C. (2010). Demonstration of the multiple matrices simply technique in establishing the psychometric characteristics of large samples. Journal of education and practice vol 2 (3).

[7] Baghaei, P. & Amrahi (2011). "Rasch Model as a construct validation tool" in Rasch Measurement Transaction, vol 22: 1, pp. 1145-1146

[8] Bond, T. G. & Fox, C. M. (2001). Applying the Rasch model: Fundamental Measurement in Human Sciences, 1st ed, Mahwah, NJ: Lawrence Erlbaum

[9] Crocker, L. & Algina, J.(1986). Introduction to Classical and Modern Test Theory. Orlando, FL: Holt, Rinehart and Winston Inc.

[10] Downing, S. M. (2003) "Item response theory: Applications of Modern Test Theory ", Medical Education, vol. 37, pp. 739-745.

[11] Embretson, S.E. & Reise, S. P.(2000). Item response theory for Psychologists. Mahwah, NJ: Lawrence-Erlbaum.

[12] Hambleton, R.K. & Swaminathan,H. (1985). Item Response Theory: Principles and Applications. Boston: Kluwer.Nijhoff.

[13] Keeves, J. P & Alagumalai, S.(1999). New approach to measurement in Keeves, J. P. & G. N. Masters, G. N. (eds.) Armsterdam : Pergamon, pp. 23-42.

[14] Linacre, J. M. (2004). "When does a gap between measures matter?" in Rasch Measurement Transaction, vol 18: 3, pp. 993

[15] Linacre, J. M. (2012). A user's guide to Winsteps Messick, S. (1993). Validity in R. L. Linn. Educational Measurement (3rd ed). Phoenix: Oryx Press, pp. 13-104.

[16] Ministry of Education (2004). Curriculum Specifications for Mathematics Nitko,A. J. (1996). Educational Assessment of Students. The wright map. 2nd. ed. Englewood Cliffs, NJ: Merrill.

[17] Odili, J. N. (2010). Modern Test Theory. An unpublished lecture note, Delta state university.

[18] Olaleye, O. O. & Aliyu, R. T. (2013). Development and Validation of Mathematics Achievement Test Items Using Item Response Theory (IRT) Models in Attaining Quality Education for National Development. A paper presented and published in the Proceedings of Mathematics Association of Nigeria (MAN) at the 50th Anniversary of the Annual National conference of MAN. Pp 82-95

[19] Opasina, O. C. (2009). Development and validation of alternative to practical Physics test using item response theory model. An unpublished Ph.D thesis, University of Ibadan.

[20] Osadebe, P. U. (2010). Construction and validation of Test Items. An unpublished lecture note, Delta state university.

[21] Reza, P. ,Baghaei, P & Ahmadi, H. S. (2011). Development and validation of an English Language Teacher competency test using Item Response Theory. The international Journal of education and psychological assessment, vol 8, 2, pp 54-68.

[22] Stocking, M. L. (1999). Item response theory in Advances in Measurement in Educational Research and Assessment, J. P. Keeves & Masters, G. N. (eds.) Armsterdam : Pergamon, pp. 55-63.

[23] Wright, B. D & Panchapakesan, N. A. (1969). "A procedure for sample-free item analysis". Educational and Psychological Measurement, vol. 29, pp. 23-48.

[24] Wright, B. D. (1999). "IRT in the 1990s: Which models work best?" in Rasch Measurement Transaction, vol 6: 1, pp. 1145-1146.