Recalibrating and Restandardizing the Standard Progressive Matrices Test Using Rasch Model

Dr. Salah Edin Farah Attallah,
Associate Professor, Special Education Dept., King Saud University

Abstract *_The Standard Progressive Matrices (SPM) test is the most widely used group IQ test in the Sudan. However, many experts criticize its construction based on the classical theory in psychological measurement. Hence, the present study aimed to investigate the extent to which the items of the SPM test fit into the one-parameter Rasch Model (RM) which is considered the base of the modern theory in measurement. The study also attempted to develop new criteria for the test to explain individuals' ability levels based on their scores on the items that fit into RM. The test was administered to two groups of participants: a calibration group (N = 1200 male and female individuals ranging in age from 6 to 25) and a standardization group (N = 4000). The PASW 18 and RUMM 2020 programs were used in the statistical treatment of data. Eight items were deleted from the SPM test because they did not fit into RM, thus leaving the test with 52 items. The modified version of the test was then subjected to validity and reliability measures. The study also developed new criteria for the test by finding t-scores and deviated IQ percentages matching the various ratings of individuals' ability.*

Key Words: *Standard Progressive Matrices Test (SPM), Rasch Model, Intelligence tests, Psychometrics proprieties.*

I. INTRODUCTION

The first version of the SPM test that was developed by John Raven and his assistants appeared in 1938. Ever since, it has been used extensively, as an important IQ group measure, in assessment and diagnosis in educational, clinical and professional contexts throughout the world. Besides, it was adapted to serve the same purposes in Arabic countries. It was standardized in the Sudan based on the classical theory in psychological and educational measurement in 1998 [1]. Ever since, it has been used extensively in the Sudan in identifying gifted individuals and in educational and psychological research. Also it has been used in many other fields that require assessment of intelligence: educational, clinical, professional and military.

However, despite its great importance, the SPM test, like other intelligence measures developed in the light of the classical theory of measurement, received many criticisms. Measures developed in the light of the classical theory of measurement are criticized for not achieving objectivity of measurement. According to advocators of recent trends in psychological measurement, measurement cannot be objective unless the results taken from the test are independent of the test itself [2] ,[3].

A group of researchers and experts [4], [5], [6], [7] summarized the drawbacks of the classical theory of measurement as follows:

1- The total test score is restricted to test items. That is, the score one obtains on a given test is dependent on the same test. Thus, if an individual achieves a score of 23 on a given test, we cannot confirm that the same individual will obtain the same score (23) on another test measuring the same variable. This indicates that an individual's score differs if different items are used.

2- Lack of linearity of measurement. Linearity of measurement means that there is a constant rate of measurement hierarchy on the measured variable continuum. That is, the scores an individual obtains can be summed as if they represented a linear measure. In classical theory, individuals' scores do not give linear measurements.

3- Measurement of more than one dimension. Physical measures are known to measure a unilateral variable. This is not the case with measured developed in the light of the classical theory.

Uniformity of test scores along the level of the measured variable. The classical theory postulates that test scores represent a linear function, i.e., the higher the individual's ability, the higher his / her score and vice versa. However it was found in some cases that low level individuals may answer items above their level correctly and vice versa.

The meaning of test items varies with time. It is known that classical test items are affected by variable environmental circumstances and non-standardized test conditions. If any of the test items is deleted, the individual's score differs in a way that is difficult to predict.

Lack of constant measurement unit since measurement points are not placed on the variable continuum linearly. That

individuals' scores are dependent on test items may result in the distance between any two successive scores being different. This results in the variability of the quantitative meaning of any specified difference across a range of test scores [8].

The characteristics of test items are affected by test takers' ability. That is, difficulty and discrimination coefficients vary according to test takers' ability. Hence, an item can be easy for given test takers and difficult for others. In case the sample is relatively homogeneous, discrimination coefficients tend to be lower compared to a heterogeneous sample [9].

The test taker's total score is affected by test items. Test takers obtain higher scores if test items are easy and vice versa. This does not give a true picture of the individual's ability. That is why the outcomes of measurement differ from test to another [10].

Comparison among test takers in the trait or ability measured is limited to the application of the same test items or a equivalent set of items to all test takers. Thus, we cannot compare ability levels if test takers answer items with different difficulty levels [11].

Test reliability is affected by the testing situation since reliability, according to this theory, is established by administering the test twice or administering an equivalent version of the test. However, the testing situation of the two applications can be somewhat different and the development of completely corresponding versions of the same test is next to impossible [9].

Uniformity of measurement error variance for all test takers regardless of the fact that the performance of some test takers can be more consistent than others' performance and that the degree of this consistency varies according to test takers' ability level or the ability level measured by the test [12].

This theory does not provide a psychological explanation of how the individual attempts to answer a test item even though such explanation is necessary if we desire to predict the characteristics of scores drawn from a given population or various populations or if we wish to design tests with psychometric characteristics that suit any given population. Additionally, the meaning of test items varies with time due to changes in environmental and testing circumstances. Changing or deleting any test item changes test takers' scores. Such change in scores is difficult to predict [5].

All characteristics of the tests developed in the light of the traditional theory such as coefficients of difficulty, discrimination and reliability depend on the characteristics of test takers and test items [13].

With the advent of the modern theory of psychological measurement known as Latent Trait Theory or Item Response Theory in the 1960s, researchers started to update classical IQ tests according to the modern theory of measurement (e.g. Chissom & Hones [14], Nenty [15], El-Korashy [16], Hernandez [17], Zimowski & Wothke [18], Abo Jarad [19], Abo Moslem [20], Jad Arrab [21], Hegazi & Bany Atta [22], Zikri [23],Zikri [24] ,Ashafey & Nour Edin [25], Atteriri [2],Alaam [5]; Nour Edin [26], Masoud [27].

Similarly, researchers have been concerned with adapting the SPM test in the light of the modern theory of measurement. Many researchers have attempted to adapt the SPM test according to RM, e.g., the study conducted by Tse and Phillipson [28] on gifted students in Hong Kong and studies concerned with heredity like the one done by van Leeuwen, van den Berg and Boomsma [29] in which they tackled intelligence heredity.

A. Statement of the Problem

Concurring with the trend of updating intelligence tests developed according to the classical theory of measurement in the light of recent theories, the present researcher sought to recalibrate and restandardize the progressive matrices test in the light of RM as this test is the most widely used in the Sudan. More specifically, the study addressed the following questions:

1- To what extent do the items of the SPM test fit into RM?

2- To what extent are the items of the SPM test difficult?

3- How valid and reliable is the SPM test after calibrated according to RM?

4- What are the criteria of the SPM test after calibrated according to RM?

Aims of the Study

1- Developing a version of the SPM test that aligns with Rasch unilateral model and have good psychometric characteristics in terms of validity and reliability.

2- Extracting the criteria that explain the individual's ability on the test.

B. Significance of the Study

The significance of the study springs from the fact that it addresses an important topic in the development of psychological and educational assessments. The adaptation of traditional assessments in the light of modern theories of measurement deserves serious research efforts.

The study is also expected to provide the mental measurement field in the Sudan with an instrument with good psychometric characteristics. The study can urge similar research attempts on the revalidation of traditional assessments based on RM and other modern theories of

measurement (e.g. item response theory) at the regional and international levels.

C. Terminology

The Standard Progressive Matrices Test

It is one of the three matrices tests (colored, advanced and normal) that was developed by the English psychologist John Raven [30]. Raven published the first version of the matrices in 1938 and pursued this effort with his students for 30 syears until his death in 1970. Raven's matrices are of the non-verbal intelligence tests. Mainly it is a group test but it can also be applied individually. It measures mental competence when administered as a timed test and mental capacity if administered untimed [31]. The test consists of 60 items divided into five groups graded based on difficulty.

Rash Model

It is one of the most important latent traits models. Latent traits is a theory that postulates the existence of one (or more) basic trait that identifies one's response to test items. It was named latent traits (or abilities in cognitive tests) because they are not subject to direct observation or measurement. Factorial analysis is known to be the best methods for identifying latent traits. RM is the only one-parameter latent trait model since objectivity requirements are secured when the model's hypotheses are met: unilateralism, independence of measurement and the parallelism of slopes that are characteristic of items. It is also referred to as one-parameter model. It was developed by George Rasch. It is concerned with locating the test item on the difficulty scale of all test items. It is also concerned with calibrating one's ability levels on a given test on the same item scale.

The underlying idea in the model is that each item holds an emotional charge that contributes with other items in the scale in forming an overall emotional charge that indicates one's direction in accordance with his/her estimation of that item based on the number of calibration categories in the scale. The model estimates the charge of each item according to the probable mathematical function adopted in the model. The alignment of items with the model is then verified [32]. RM is based on the interaction between individuals' traits and difficulty of items. The results of interaction are represented in observable responses through which the calibration of items and individuals' estimations can be identified Alaam [13], Kathem [33].

D. Calibration

Calibration means estimating the difficulty of items and assessing their scores' alignment with the used model, and making use of this in estimating individuals' traits [13].

E. Standardization

Standardization refers to identifying test conditions accurately, so one way is used in applying the test, scoring it and interpreting obtained scores. This entails controlling for all variables that can affect results, so results can be used to attribute differences among individuals to individual's traits and not other variables [34]. Standardization involves the scale's objectivity, criteria, means, reliability and validity.

II. Review of Literature

Gallini [35] used Rash's one-parameter model in analyzing the items of the progressive matrices test. The study aimed to test the extent to which the observed performance fit with the expected performance on the matrices test using Rash's one-parameter model. The sample consisted of 151 seventh graders. The obtained data were then treated statistically using the Bical program. After the first analysis, 4 subjects were eliminated based on the level of significance (+ or -2). This way, subjects who were not serious were eliminated. Four items were eliminated based on significance (+ or -2.5). This way, unsuitable items were eliminated. The difficulty of the items ranged between -16 to 4.4 logits. Results revealed low homogeneity among test items calibrated according to Rash model where dual correlation coefficients ranged between .04 and .57.

The study conducted by Saccuzzo and Johnson [36] aimed to convert raw scores on the normal and advanced forms of Raven's matrices, so they are graded along a common calibration scale using parity with equal percentiles to obtain one common scale that suits the wide range of mental ability. Participants were 261 university students. The study found equivalent scores on the normal and advanced forms of the matrices test, which facilitates the comparison of scores.

Green and Kluever [37] recalibrated the progressive matrices test (the normal and colored forms) to identify the effect of item elements on the relative difficulty of Raven's items. The sample consisted of 269 students and 151 academically gifted seventh graders whose ages ranged between 9.2 and 11.8 years. The researchers used the multiple regression equation to analyze the data represented in the characteristics composing items regardless of the mental processes required for performance after calibrating test items with RM. Results indicated that the regression coefficient of the colored form (.9) is higher than its counterpart of the normal form (.69). It was also found the components of the test items of the progressive matrices test affect the relative difficulty of test difficulty estimations which were obtained through analyzing the test using RM.

Styles and Andrich [38] conducted a study to calibrate Raven's test using RM and merge the normal and colored forms of the test in one measure. The new measure was

applied in two modes: the paper-and-pencil mode and computer-adaptive mode. The results of the two modes were then compared in terms of item difficulty estimations and statistics of appropriateness. The sores on the normal form of the test were transformed to scores on the advanced form in both traditional and programmed modes. The test was administered to two samples: 909 students for the traditional mode and 190 for the programmed mode. Each sample was divided into age groups. The resulting instrument was an adapted form of Raven's progressive matrices test with 96 items (60 for normal matrices and 36 for advanced matrices). This adapted test was administered in a traditional mode and a programmed mode to three different groups. The final form of the programmed test consisted of 89 items including 5 preparatory items and 2 items that are lower in difficulty than students' level.

Kubinger, Formann and Farakas [39] examined the psychometric characteristics of Raven's SPM test using RM. The test was given to 527 students whose ages ranged between 8 and 14 years. Only 17 items fitted into RM, whereas 60 items did not. When given to another sample, the 17 items did not fit into the model. The researchers concluded that the relationships among items were not independent of the context in which the items are presented.

In the study conducted by van der Ven and Ellis (2000 cited in Eed, 2005[40] the SPM test was administered to 901 students between 12 and 15 years of age to analyze the 5 sub-components separately. Three of the 5 subtests (A, C and D) fitted into RM, while the other two (B and E) did not.

Attantawy [41] used RM to adapt Raven's progressive matrices. The sample insisted of 1411 elementary and preparatory school students whose ages ranged between 6 and 13 years. A clear match between the order of the items in the final form of the test and the original form was found. Eleven items did not fit into the model. The final form included 49 items with a reliability coefficient (Kuder-Richardson Equation) of .985.

Eed [40] examined the structure of Raven's advanced progressive matrices test and its shortened form through factor analysis and RM. Two groups were used: one group was given the complete test (N = 500 students: 320 females and 180 males) and the other was given the shortened test (N = 640 students: 416 females and 224 males). Analysis based on RM demonstrated that the complete test did not fit into the mode since approximately half of the items needed to be eliminated. It was also found that the test was not unilateral. Fifteen students of those who took the short form were eliminated as they answered all items correctly. The shortened form proved to have three various factors, i.e. it was not unilateral.

Vigneau and Bors [42] tackled the issue of dimensionality (unidimensionality versus multidimensionality) in the advanced progressive matrices test. They examined dimensionality in the original form of the test (SET II) on a sample including 506 examinees and the shortened form on a sample including 644 examinees using the Principal Component Method and RM. Although results of factor analysis were equivalent, results of RM were more vigorous, thus indicating that the two forms were carefully developed to be multidimensional. Comparison of the two forms revealed an effect of performance context, which makes difficult developing adapted tests form them.

Al-Qafas [43] sought to propose a method to rate individuals' abilities by the scale after calibrated. This method is based on using the difficulty of calibrated items in setting weights to items, so an item's score is not equal to another item's score unless the two items are equal in difficulty. The calibration sample consisted of 312 students ranging in age between 7 and 16 years. The progressive matrices test was administered on the study's basic sample (N = 2099) after calibrated according to RM. The data of the calibration sample was analyzed to develop the scale. The produced sale was then given to the basic sample and the total raw score of each student was converted to an ability using the ability table that calibration produced. Item weights were then computed using item difficulty that calibration produced. The examinee's ability was estimated using the proposed method (the sum of the weights of the items answered correctly).

Comments on Reviewed Studies

1. The test has captured researchers' interest throughout the past 3 decades to adapt based on the modern theory of measurement.
2. Some studies were concerned with examining dimensionality in the test, calibrating two tests of the matrices in one test, and suggesting new methods for estimating difficulty or rating individuals' abilities.
3. Studies found that the SPM test in its original form did not fit into Rash model. For the test to fit into the model, from 4 to 11 items had to be eliminated and items had to be reordered. Only one study found 17 fit items.
4. The majority of the studies were concerned with calibrating the test according to Rasch one-parameter model. No study addressed calibration, standardization and extraction of new criteria for the test. What makes the present study of potential value is that it would attempt to extract new criteria for the test after calibrating it according to RM. The new criteria would be estimated by deviated intelligence percentages and t-sore. The present study is also distinguished in that it would use a large sample of students from different age groups (from 6 to 25 years). Finally, the study would use the RUMM

2020 program that is void of the drawbacks in previous statistical programs.

III. METHOD AND PROCEDURES

A. PARTICIPANTS

Two groups were used: a calibration group that would be used for calibrating test items and a standardization group that would be used to extract test criteria.

A. The Calibration Group

The test was given to 1200 male and female students (60 for every age group of each gender) ranging in age from 6 to 25 years. All students were from Khartoum.

B. The Standardization Group

This group consisted of 400 male and female students (200 for every age group of each gender) whose ages ranged between 6 and 25 years. All students were from Khartoum.

B. PROCEDURES

- The test was given to participants according to the instructions included in the test guide,
- Answers were coded and fed into the PASW 18 statistical program,
- Data was treated and fed into the RUMM 2020 program to analyze it using Rasch one-parameter probable logarithmic model and identify the characteristics of the items and individuals.
- The criteria explaining individuals' different levels were computed through the standardization group.
- T-scores and deviated intelligence percentages (IQ index) matching ability estimates according to Wechsler equation were computed.

IV. RESULTS

First: to answer the first research question: To what extent do the items of the SPM test fit into RM?

The researcher analyzed the 60 items making the test to examine the extent to which items fit into RM using the RUMM 2020 program. What follows are the results in terms of:

- Elimination of whole and zero data from the analysis matrix. This included:
- Eliminating participants obtaining the full test score since the ability of such individuals exceed the range covered by the test. No such participants were found. Thus, the number of participants remained unchanged (i.e. 1200).
- Eliminating participants obtaining no score since the ability of such individuals is lower than the range covered by the test. No such participants were found.
- Eliminating items to which all participants responded correctly. This resulted in eliminating the first item which is presented to examinees as an illustrative example of how to answer items. This left the test with 59 items.
- Eliminating items to which all participants did not respond correctly as items of this nature have no discriminating power. No items were eliminated for this reason.
- Eliminating participants who do not fit into the model

After completing the previous step, analysis began to eliminate participants who do not fit into the model, i.e., participants unfit for the calibration process. This was achieved in the light of the following criteria:

- Eliminating participants whose appropriateness value was less than -2. This demonstrates the similarity of the rating obtained by such participants, i.e., their responses are not valid.
- Eliminating participants whose appropriateness value was more than +2. Such respondents exceed the statistically acceptable limit.
- The previous two steps resulted in eliminating 30 participants. Thus, 1170 participants remained in the sample.
- Eliminating items that do not fit into the model:

Data was reanalyzed to eliminate items that do not fit into the model, i.e., to eliminate items having defects that make them unsuitable for calibrating the measured variable. This was done according to the following criteria:

- Eliminating items with appropriateness value less than -2.5 as this means that such items are not independent of other test items or that they measure another variable that is very similar to the measured variable.
- Eliminating items with appropriateness value more than +2.5 as this signifies that there is a defect in item construction or that the items measures another variable.

This analysis was performed several times on different samples. Based on this, 7 items were eliminated: 31-c7, 34-c10, 40-d4, 42-d6, 50-E2, 54-E6, and 60-E12. This procedure left the test with 52 items that fitted into RM. Then the initial characteristics of the test and the participants were extracted as shown in table 1:

Table 1
Summary of the Analyses of the SPM test

Items & Individuals	Statistical Indices	Logit	Appropriateness of Residuals
Items	Mean	0.000	- 0.150
	Standard Deviation	2.078	1.180
	Skewness	-	0.155
	Kurtosis	-	-0.724
	Correlation	-	0.250
Individuals	Mean	0.730	-0.201
	Standard Deviation	1.502	0.808
	Skewness	-	1.169
	Kurtosis	-	1.469
	Correlation	-	-0.110

As listed in Table 1, the total value of Chi Square for items was 516.529 with degree of freedom of 312 and the probability score of Chi Square was .0000. The power of test appropriateness was high.

Second: to answer the second research question: To what extent are the items of the SPM test difficult?

After the procedures in 1, 2 and 3, the researcher reached the final calibration of the SPM test based on item difficulty (see table 2) and then set the new calibration of test items as demonstrated in table 3:

Table 2
The characteristics of SPM test
Items in terms of difficulty and appropriateness

Item no. before calibration	Logit	Minf	Normal Error	Appropriateness of Residuals	χ^2	probability
A 2	-5.132	24.34	0.513	-0.706	0.867	0.990144
A 3	-2.882	35.59	0.222	0.771	34.230	0.000006
A 4	-3.129	34.355	0.239	0.705	16.363	0.011933
A 5	-3.468	32.66	0.266	-0.029	20.339	0.002410
A 6	-3.328	33.36	0.254	1.242	14.653	0.023133
A 7	-1.282	43.59	0.151	-1.900	5.746	0.452183
A 8	-1.777	41.12	0.168	-0.328	5.593	0.470285
A 9	-1.831	40.85	0.170	-0.548	1.957	0.923596
A 10	-0.644	46.78	0.135	-0.171	10.073	0.121632
A 11	0.237	51.19	0.121	0.703	8.562	0.199746
A 12	1.056	55.28	0.116	2.314	8.894	0.179642
B 1	-3.705	31.48	0.289	0.202	9.104	0.167811
B 2	-2.992	35.04	0.229	-1.122	3.022	0.806128
B 3	-1.697	41.52	0.165	-0.562	9.614	0.141877
B 4	-0.694	46.53	0.136	-1.113	7.621	0.267218
B 5	-0.330	48.35	0.129	-1.865	11.130	0.084437
B 6	-0.077	49.62	0.125	-1.045	6.478	0.371862

Table (2) continue

B 7	0.406	52.03	0.119	-0.348	4.305	0.635456
B 8	1.128	55.64	0.116	-1.171	5.173	0.521882
B 9	0.628	53.14	0.117	-2.081	7.931	0.243192
B 10	0.039	50.20	0.123	-0.903	4.295	0.636775
B 11	0.630	53.15	0.117	1.268	6.796	0.340112
B 12	1.338	56.69	0.117	0.647	9.559	0.144497
C 1	-1.890	40.55	0.172	-0.567	1.705	0.944751
C 2	-1.158	44.21	0.148	-0.340	6.315	0.388806
C 3	-1.061	44.70	0.145	-2.312	10.316	0.111967
C 4	0.038	50.19	0.123	-0.595	4.679	0.585566
C 5	-0.254	48.73	0.128	-0.958	7.087	0.312850
C 6	1.021	55.11	0.116	1.007	3.108	0.795117
C 8	1.281	56.41	0.116	-0.906	3.205	0.782731
C 9	0.488	52.44	0.118	-0.372	4.740	0.577615
C 11	1.666	58.33	0.119	0.485	5.777	0.448685
C 12	3.540	67.70	0.173	1.740	44.740	0.000000
D 1	-1.906	40.47	0.172	-1.384	8.600	0.197361
D 2	-1.233	43.84	0.150	-2.217	9.627	0.141287
D 3	-0.965	45.18	0.143	-0.848	12.471	0.052246
D 5	-1.058	44.71	0.145	-1.834	12.978	0.043384
D 7	0.183	50.92	0.121	-1.881	13.160	0.040571
D 8	0.354	51.77	0.120	-0.574	9.775	0.134475
D 9	0.359	51.80	0.120	1.521	5.080	0.533581
D 10	0.911	54.56	0.116	-0.177	13.209	0.039835
D 11	2.202	61.01	0.126	0.028	10.187	0.116990
D 12	3.489	67.44	0.170	0.063	9.287	0.158075
E 1	0.223	51.12	0.121	0.706	3.062	0.801056
E 3	1.111	55.56	0.116	-0.295	5.479	0.483955
E 4	2.169	60.85	0.126	1.979	10.644	0.100013
E 5	2.198	60.99	0.126	2.205	8.703	0.191000
E 7	1.914	59.57	0.122	1.624	16.673	0.010565
E 8	3.511	67.56	0.171	-0.066	5.742	0.452717
E 9	3.030	65.15	0.150	0.697	8.183	0.224997
E 10	3.414	67.07	0.166	0.582	27.706	0.000107
E 11	3.931	69.66	0.197	0.919	21.987	0.001219

Degrees of freedom 1 and 2 are 402.12 and 6 respectively

Table 3
The new calibration of SPM test

Item no. before calibration	Item no. after calibration	Item difficulty (Logit)	Item no. before calibration	Item no. after calibration	Item difficulty (Logit)
0.223	27	H 1	-5.132	1	A 2
0.237	28	A 11	-3.705	2	B 1
0.354	29	D 8	-3.468	3	A 5
0.359	30	D 9	-3.328	4	A 6
0.406	31	B 7	-3.129	5	A 4
0.488	32	C 9	-2.992	6	B 2
0.628	33	B 9	-2.882	7	A 3
0.630	34	B 11	-1.906	8	D 1
0.911	35	D 10	-1.890	9	C 1
1.056	36	A 12	-1.831	10	A 9
1.021	37	C 6	-1.777	11	A 8
1.111	38	H 3	-1.697	12	B 3
1.128	39	B 8	-1.282	13	A 7
1.666	40	C 11	-1.233	14	D 2
1.281	41	C 8	-1.158	15	C 2
1.338	42	B 12	-1.061	16	C 3
1.914	43	H 7	-1.058	17	D 5
2.169	44	H 4	-0.965	18	D 3
2.198	45	H 5	-0.694	19	B 4
2.202	46	D 11	-0.644	20	A 10
3.030	47	H 9	-0.330	21	B 5
3.414	48	H 10	-0.254	22	C 5
3.489	49	D 12	-0.077	23	B 6
3.511	50	H 8	0.038	24	C 4
3.540	51	C 12	0.039	25	B 10
3.931	52	H 11	0.183	26	D 7

Third: to answer the third research question: How valid and reliable is the SPM test after calibrated according to RM?

A. Validity of Calibration

The calibration of items measuring the same trait on a common scale using RM means that they measure one variable. This unidimensionality of measurement in RM secures calibration validity of items in measuring the intended variable. It also ascertains calibration validity of individuals' abilities on the variable continuum which is based on the validity of their responses to items [8]. Unidimensionality is met if examinees and items fit into the model based on the criteria of the investigator.

B. Factorial Validity

The researcher conducted factor analysis on the calibration sample using the Maximum Likelihood Method. It was found that KMO equivalence coefficient, Bartlett's test of Spherity and degrees of freedom were .941, 551113.118 and 1326 respectively (p = .000). Analysis also revealed that item communalities were high (they ranged between .423 and .713). The most important finding was that items loaded on six factors and that the latent root of the first factor was .713,

explaining 25.912 of total variance. This means that the test is unidimensional. The test appropriateness quality was high (χ^2 = 4903.379 with degrees of freedom of 768 and significance level of .000). That is, the test proved valid.

C. TEST RELIABILITY

Reliability of Calibration

The calibration of test items on a common scale according to RM after eliminating items and individuals that do not fit into the model means that the model's conditions including independence of measurement are met. This ascertains that difficulty and ability ratings are reliable and that they are not affected by changing the set of variable drawn from the original calibration scale or changing the participants taking the test.

Reliability Coefficient established by the computer program

The RUMM 2020 program establishes reliability according to the traditional theory of measurement. The test yielded an alpha reliability coefficient of .93. Person Separation Index was .928.

Fourth: to answer the fourth research question: What are the criteria of the SPM test after calibrated according to RM?

The researcher established the criteria that explain an individual's ability on the test by extracting t-sores and deviated intelligence percentages according to Wechsler equation. RM only calibrates test items based on whether or not they fit into it and estimates individuals' levels on the test by logit and minf units. Thus, the identification of test criteria was achieved by using group referenced criteria matching ability ratings on the test. Test criteria were established as follows:

- Computing the total raw score of each participant in the standardization sample on the final form of the test (52 items),
- Converting total raw scores of all participants into matching ability ratings using the table for identifying probable matching ability ratings for each possible total score,
- Computing mean and standard deviation of individuals' ability on the test through the minf unit by converting the logit into minf based on the linear conversion equation and
- Computing the criteria of t-scores and deviated intelligence percentages for each of the ability ratings as expressed in minf units. The following table shows the results of these procedures:

Table 4
Matching Ratings for Each
Possible Total Raw Score on the SPM Test

Total raw score	Matching ability				Total raw score	Matching ability			
	Logit	Minf	t-score	Special education score		Logit	Minf	t-score	Special education score
0	-5.477	22.615	- 4.77	17.845	**27**	0.186	50.93	51.86	102.79
1	-4.610	26.95	3.9	30.85	**28**	0.291	51.455	52.91	104.37
2	-4.000	30	10	40	**29**	0.395	51.975	53.95	105.93
3	-3.750	31.25	12,5	43.75	**30**	0.499	52.495	54.99	107.49
4	-3.229	33.855	17.71	51.57	**31**	0.604	53.025	56.04	109.06
5	-2.943	35.285	20.57	55.86	**32**	0.710	53.55	57.10	110.65
6	-2.794	36.03	22.06	58.09	**33**	0.817	54.085	58.17	112.26
7	-2,471	37.65	25.30	62.95	**34**	0.925	54.625	59.25	113.88
8	-2.269	38.655	27.31	65.97	**35**	1.053	55.265	60.53	115.80
9	-2.082	39.59	27.31	65.97	**36**	1.148	55.74	61.48	117.22
10	-1.909	40.455	29.18	68.77	**37**	1.263	56.315	62.63	118.95
11	-1.746	41.27	30.91	71.36	**38**	1.381	56.905	63.81	120.72
12	-1.593	42.035	32.54	73.81	**39**	1.263	56.315	62.63	118.95
13	-1.447	42.765	34.07	76.11	**40**	1.631	58.155	66.31	124.47

Table (4) continue

14	-1.307	43.465	35.53	78.30	**41**	1.763	58.815	67.63	126.45
15	-1.174	44.13	36.93	80.40	**42**	1.903	59.515	69.03	128.55
16	-1.045	44.775	38.26	82.39	**43**	2.051	60.26	70.51	130.76
17	-0.921	45.395	39.55	84.33	**44**	2.210	61.05	72.10	133.15
18	-0.800	46	40.79	86.19	**45**	2.382	61.91	73.82	135.73
19	-0.682	46.59	42.00	88.00	**46**	2.572	62.86	75.72	138.58
20	-0.568	47.16	43.18	89.77	**47**	2.785	63.93	77.85	141.78
21	-0.455	47.725	44.32	91.48	**48**	3.031	65.16	80.31	145.47
22	-0.345	48.275	96.55	94.83	**49**	3.328	66.64	83.28	149.92
23	-0.237	48.815	97.63	96.45	**50**	3.709	68.55	87.09	155.64
24	-0.130	49.35	98.70	98.05	**51**	4.264	71.32	92.64	163.96
25	-0.024	49.88	99.76	99.64	**52**	5.073	75.37	100.73	176.10
26	0.082	50.41	100.82	101.23					

V. Conclusions and Recommendations

The present study aimed to recalibrate and restandardize the SPM test using Rasch one-parameter model, and to establish different criteria explaining individuals' ability levels. Eight items were eliminated from the test for not fitting into RM. The recalibrated test then included 52 items. The study also established criteria for the test by obtaining t-scores and deviated intelligence percentages matching individuals' various ability ratings.

The study therefore made use of the linearity of measurement that characterizes RM where there is one measurement unit for item difficulty and the examinee's ability, the Logit unit. The Logit was converted in the present stuffy into the Minf unit, deviated intelligence percentages and t-scores. The eliminated items were relatively few (8 items). This finding is similar to the findings of Gallini [35] and Attantawy [41]. The study also found a difference in the order of test items before and after calibration. The order of test items after calibration is more logical.

The study ascertains the assumption that traditional IQ test demonstrate high psychometric characteristics when used with latent trait models. In the respect, the present study concurs with the studies of Abo Jarad [19], Abo Moslem [20], Jad Arrab [21], Zekri [23], Zekri [24], Attweiri [2], Ashafey and Nour Edin [25], Alaam [5], Massoud [27], Chissom and Hones [14], El-Korashy [16], Zimowski and Wothke [18], Nenty [15], and Hernandez [17]. The results of the calibration of the SPM test show that IQ tests developed in the light of the latent trait theory have high psychometric characteristics. This is reflected on the accuracy of the criteria established for such tests.

The results ascertain the importance of using the latent trait models in analyzing scores of IQ tests. They also ascertain that statistical analyses should not be based only on raw scores. Latent trait models and other modern models of psychological measurement yield more accurate results. This leads to better decision making. Finally, the study ascertained the use of RM in developing group IQ tests.

Based on the results reached, the following recommendations are offered:

- Using the adapted form of the SPM test that the present study came up with in IQ measurements in Khartoum and using the new standard scores in interpreting individuals' score.
- Using RM in adapting other PM tests and other group IQ tests in the Sudan.

References

[1] Alkhateeb, Mohammed & Almotawakel, Meheid. (2002). A pilot study of the psychometric characteristics of the standard progressive matrices test. *Psychological Studies*, 1, 89-102.

[2] Atteriri, Abdurahman. (1996). The psychometric characteristics of preparatory IQ test according to RM. *Psychological Studies*, 4, 457-473.

[3] Alaam, Salah Edin. (2005). *One-dimension and multi-dimensions item response models and their applications in psychological and educational measurement*. First Edition, Amman, Dar Alfekr Alarabi.

[4] Addardir, Abdulmonem. (2004). Calibrating high IQ test using Rasch one-parameter model. In Abdulmonem Addardir (Ed.). *Contemporary Studies in Educational Psychology*, 12-94. Cairo: Alam Alkotob.

[5] Alaam, Salah Edin. (1985). Analyzing data of mental tests according to Rasch Logarithmic Probable Model: an experimental study. *The Arabic Journal of Human Sciences*, 5(17), 100-122.

[6] Kathem, Amenah. (1988). *A critical theoretical study of objective measurementp0 of behavior: Rasch Model*, Kuwaiti Institution for Scientific Development.

[7] Murad, Salah & Ashafey, Mohammed. (1988). The effect of sample size on the accuracy and efficiency of joining two tests in one calibration. *Journal of Psychological and Educational Research*, 2, 51-97.

[8] Kathem, Amenah. (1996). Latent traits models. In Anwar Asharqawy, Solyman Asheikh, Aminah Kathem & Nadia Abdusalam, 1996 (EDS.), *Temporary trends in psychological and educational measurement and evaluation*, 281-430, Cairo: Alanglo Almasriyah.

[9] Hambleton, R. & Swaminathan, H. (1989). *Item response theory principles and applications*. Boston, Kluwer Nijhoff Publishing.

[10] Alaam, Salah Edin. (2000). *Educational and psychological measurement and evaluation: its bases, applications and contemporary trends*. Cairo, Dar Alfekr Alarabi.

[11] Abdelmaseh, Emad (1991). Using Rasch Logarithmic One-Parameter Model in analyzing norm referenced bipolar cognitive tests: an experimental study. *Journal of Educational and Psychological Research*, 4, 443-457.

[12] Abo Hashem, Assayed. (2006). A comparison between the traditional theory and RM in selecting items of Study Approaches Scale among university students. *Faculty of Education Journal, Zagazig University*, 52, 17-70.

[13] Alaam, Salah Edin. (1987). A critical comparative study of latent traits models and classical models in psychological and educational measurement. *The Arabic Journal of Human Sciences*, 27, 18-44.

[14] Chissom, B., & Hoenes, R. (1976).A comparison of the ability of the D-48 and IPAT culture fair intelligence test to predict SAR achievement test scores for 8th and 9th grade student. *Educational and Psychological Measurement, 36*, 561-564.

[15] Nenty, H. J. (1986). *Cross -cultural bias analysis of Cattell's culture fair intelligence test* .Paper presented at the annual meeting of the American Educational Research Association. 70th. San Francisco, CA. April 16-20.

[16] El-Korashy, A. (1995). Applying the RM to the selection of items for a mental ability test. *Educational and Psychological Measurement, 55(5)*, 753-763.

[17] Hernandez, Royce. (2009). Comparison of the Item Discrimination and Item Difficulty of the Quick-Mental Aptitude Test using CTT and IRT Methods. *The International Journal of Educational and Psychological Assessment, 1(1)*, 12-18.

[18] Zimowski, M. & *Wothke*, W. (1987). *Purification of spatial tests and Reasoning components in spatial tests*. Paper presented at the Annual Meeting of the American Educational Research association, (Washington, Dc, April 20-24).

[19] Abo Jarad, Hamdy. (2008). Using Rasch Model in developing Form A of Cattle's Third IQ test. *The Islamic University Journal, the Human Studies Series*, 2(16), 555-583.

[20] Abo Moslem, Maysaa. (2010). Equating the two versions of Tony's nonverbal IQ test using different methods of the equation in the light of some factors affecting its results. *The Egyptian Journal of Psychological Studies*, 66, 370-411.

[21] Jad Arrab, Hesham. (1999). *Developing Cattle's test using latent trait models and the effect of this on the test's ability to predict scholastic achievement*. Unpublished M.A thesis, Faculty of Education, Mansoura University.

[22] Hegazi, Taghreed & Bani Atta, Zayed. (2010). Investigating the extent to which the responses to the Jordanian adaptation of Ottis-Lennon's mental ability test fit into the modern theory of measurement. *Journal of Al-Shareqa University for Human and Social Studies*, 7(2), 1-28.

[23]Zekri, Ali. (2009). *The psychometric characteristics of Ottis-Lennon mental ability test according to classical measurement and RM among intermediate students*. Unpublished Ph. D. Dissertation, Om AlQora University: KSA.

[24] Zekri, Ali. (2011). Developing and calibrating version J of Ottis-Lennon mental ability test using RM. *Faulty of Education Journal, Mansoura University*, 75(2), 64-119.

[25] Ashafey, Mohammed & Nour Edin, Ameen. (2007). Using the partial mathematical logarithmic rating model in developing primary metal abilities test on sample from the Saudi environment. *Faulty of Education Journal, Zagazig University*, 56, 245-345.

[26] Nour Edin, Ameen. (1995). *Some psychometric characteristics of Stanford-Bient modified scale among samples of preschoolers*. Unpublished M.A thesis, Faculty of Education, Ain Shams University.

[27] Masoud, Waleed. (2004). *A psychometric study for developing drawing man test using Rasch Model*. Unpublished M.A thesis, Ain Shams University, Egypt.

[28] Phillipson, S. & Tse, A. (2007). Discovering patterns of achievement in Hong Kong students: An application of the Rasch measurement model. *High Ability Studies, 18(2)*, 173–190.

[29] van Leeuwen, M., van den Berg, S. & Boomsma, D. (2008). A Twin-Family Study of General IQ. *Learning and Individual Differences, 18 (1)*, 76-88.

[30]Raven, J. (1958). *Standard Progressive Matrices A, B, C, D, and E*. London: H. K. Lewis and Co. Ltd.

[31] Raven, J. (1952). *Human Nature*. London: H. K .Lewis Co.

[32] Odda, Ahmed. (1992). Conformity between RM and classical indices in selecting items of a 7-Likert scale of attitude. *Faulty of Education Journal, Emirates University*, 8, 153-179.

[33] Kathem, Amenah. (2000). Contemporary trends in question banks. *In educational bases of university teacher preparation*. Third Edition, 321-342. Cairo: Ain Shams University.

[34] Karaja, Abdulqader. (2001). *Measurement and evaluation in psychology: a new perspective*. The Second Edition. Amman, Dar Alyazouri.

[35] Gallini, J. (1983). A Rasch Analysis of Raven Item Data. *Journal of Experimental Education, 52 (1)*, 27-32.

[36] Saccuzzo, D. & Johnson, N. (1988). Equating the standard and advanced forms of the Raven progressive matrices. *Psychological Assessment, 7 (2)*, 183-194.

[37] Green, K. & Kluever, R. (1991). *Component Identification Item Difficulty of Ravens Matrices Items*. Paper presented at the Annual meeting of the national council on measurements in education (Chicago, IL, April 4- 6, 1991).

[38] Styles, I. & Andrich, D.(1993). Linking the Standard and Advanced Forms of the Raven's Progressive Matrices in both the Pencil-and-Paper and Computer-Adaptive-Testing Formats. *Educational and Psychological Measurement, 53 (4)*, 905-25.

[39] Kubinger, K., Formann, A., & Farakas, M. (1998). Psychometric shortcomings of Ravens Standard Progressive Matrices, in particular for computerized testing. *Revue Europeenne de Psychologie bappliqué, 41*, 295-300.

[40] Eed, Khaledah. (2005). Examining the structure of Raven's advanced progressive matrices test and its shortened form using factor analysis and Rasch Model. *Journal of Psychological and Educational Studies*, 3, 256-283.

[41] Attantawy, Mona. (2004). *A psychometric study on the development of Raven's progressive matrices test according to Rash Model*. Unpublished M.A thesis, Girls' College, Ain Shams University.

[42] Vigneau, F. & Bors, D. (2005). Items in Context: Assessing the Dimensionality of Raven's Advanced Progressive Matrices. *Educational and Psychological Measurement, 65 (1)*, 109-123.

[43] Al-Qafas, Waleed. (2006). Estimating the abilities of individuals with abnormal and normal response models according to RM using W-Ratio of items according to their difficulty levels. *Psychological Studies*, 16(1), 137-160.