Full paper

Sains
Humanika

To Determine The Effect Of Raters On Science Process Skills Performance Assessment Among Primary School Students.

Gopal Krishnan Govindasamy, Mohd Ali Samsudin, Rozaini Abu Bakar*

School of Education Studies, University Science Malaysia, 11800 USM ,Pulau Pinang, Malaysia

*Corresponding author: rozek_98@hotmail.com

Abstract

This study is done to know whether the judges' ratings results in the same decisions for candidates of the same ability and to show the relative severity of the different judges. The Many-facets Rasch Measuring Model (MFRM) is used in this study to determine the validity and reliability of raters' severity and leniency. The research involves five raters and fifty examinees from standard five students of a primary school in Butterworth. There are five science process skills evaluated that is the observation, classifying, measuring, relationship between space and time and experimenting (identifying variables, make hypothesis, presenting the report). The results of this study suggest that raters of science process skills performance such as those in the PEKA can be trained to rate appropriately and consistently, and that under a system of double marking, assigning different raters to different test takers does not pose a threat to the validity of scores, and that tests are valid, reliable, and fair in that regard.

Keywords: Many-facet Rasch measurement ; rater effect; science process skill.

Abstrak

Kajian ini bertujuan untuk mengetahui sama ada penilaian yang dilakukan oleh guru-guru penilai memberi keputusan yang sama terhadap calon-calon yang mempunyai keupayaan yang sama dan untuk menunjukkan tahap ketegasan antara guru-guru penilai yang berbeza. Model Pengukuran Rasch Pelbagai-Faset (Many-facets Rasch Measuring Model) digunakan dalam kajian ini untuk menentukan kesahan dan kebolehpercayaan dalam ketegasan dan kelonggaran pentaksiran oleh guru-guru penilai. Penyelidikan ini melibatkan 5 orang guru penilai dan 50 orang murid Tahun 5 dari salah sebuah sekolah rendah di daerah Butterworth, Pulau Pinang. Terdapat lima kemahiran proses sains yang dinilai iaitu pemerhatian, pengelasan, pengukuran, hubungan antara ruang dan masa dan eksperimen (mengenal pasti pembolehubah, membuat hipotesis, menyampaikan laporan). Hasil dapatan kajian menunjukkan bahawa prestasi guru-guru penilai dalam kemahiran proses sains seperti PEKA, boleh dilatih dalam membuat penilaian dengan cara yang betul dan konsisten. Dalam sistem pemarkahan berganda, penetapan guru penilai yang berbeza untuk pengambil ujian yang berbeza tidak menimbulkan ancaman kepada kesahan skor manakala ujian adalah sah, boleh dipercayai, dan adil.

Kata kunci: Model Pengukuran Rasch Pelbagai-Faset; kesan penilai; kemahiran proses sains.

1.0 INTRODUCTION

Several researchers over the past years have suggested that the laboratory is not only a unique resource for teaching and learning, but also a unique point for observing students' idea and for assessing this new area. Students' performance, understandings, and perceptions of the science laboratory learning situation should be assessed with the use of instrument and strategies that are more closely aligned with the unique activities and goals for learning associated with the school laboratory. Grobman (2007) identified a major problem in assessing 'hands on' performance that persists to this day: "With few exceptions, evaluation has depended on written testing. There has been little testing which requires actual performance in a real situation or in a simulated situation which approaches reality to determine not whether a student can identify a correct response, but whether he can perform an experiment or projects"

Science is a process involves an integration of knowledge, skills and attitudes to understand the environment. So, teaching of science includes teaching of science process skills. Scientific process skills are known as procedural skills, experimental science, and scientific inquiry abilities (Harlen, 1999). In thinking and working scientifically, scientists use their understanding of evidence to answer questions and solve problems in such a way that it produces evidence (Skamp, 1998). Students are exposed to experiments which require the skills of observing, finding, predicting, hypothesizing, evaluating and interpreting data. Therefore, this leads to another area, assessing the science process skills through "hands on", accumulating folios on a particular topic, question-answer session and other suitable methods. This is called the performance based assessment.

Performance based assessment has been defined as "the execution of some task or process which has to be assessed through actual demonstration, that is, a productive activity (Wiggins, 1993). Moon and Callahan (2001) report that performance assessments have become very popular classroom assessments for the past twenty years. The popularity of performance assessment has given rise to number of studies done surrounding this type of assessment including its reliability and validity.

In Malaysia, Science Practical Assessment (PEKA) is a school based assessment that is implemented at school level as part of teaching and learning process. A guide is formatted by the Malaysian Examination Board (Lembaga Peperiksaan Malaysia) which contains information on the objectives, characteristics and organization of PEKA. The guide is well prepared by the board which went through many phases of development with the involvement of intelligence from the think tank groups, higher learning institution and experience science teachers. It also went through many pilot testing before it was finally accepted as the official guide to assess the students' performance assessment throughout the country. The assessment of PEKA is carried out as part of teaching and learning process. Teachers can assess either one construct or skill or several construct or skills to a small group of pupils or the whole class. Pupils who have not mastered any assessed constructs are able to repeat it in another assignment and should be given adequate chances to master the required skills before assessment is made.

■2.0 PERFORMANCE ASSESSMENT, RATER EFFECTS AND MANY-FACET RASCH MEASUREMENT

Performance assessments require test takers to perform actual tasks that are similar or relevant to the knowledge, skill, or ability being measured, and success or failure on the tasks are typically judged by human raters as done in the Practical Science Assessment in Primary Schools better known as Penilaian Kerja Amali (PEKA) (Kane, Crooks, & Cohen, 1999). There are problems connected with the use of performance assessments. The first has to do with the practical limits of doing tasks because performance assessments tend to require more time, examinees are typically tested on one or two tasks and scored on the basis of these limited samples. It is unclear whether performance on a small number of tasks is sufficient for representing domains as apparently complex and multi-faceted as other learning skills. Thus, there is the risk of *construct underrepresentation* (Messick, 1996) in assessments of this kind, and their use raises questions about generalization

Scoring is also a more difficult work in performance assessment. Scoring of performance assessments usually require the judgement of a human raters. The introduction of subjectivity into the scoring process can increase *construct-irrelevant variance* (Messick, 1996). Raters of performance assessments come from many different backgrounds, factors they actually consider, beliefs they bring to the rating task, which is not clear and misunderstood, and threatens to assume the ratings they give invalid

The introduction of performance assessment not only brought with it promises of greater validity but also greater risks of unwanted variability (Linacre, 1989; McNamara, 1996; Wilson & Case, 2000). Performance assessment, unlike the traditional fixed-response assessment, has features that are peculiar to its assessment setting – the task choice, the task processing conditions, the raters, the rating scale and the rating procedures that involve subjectivity of human judgment. (McNamara, 1996; Upshur & Turner, 1999).

The rater severity is the most widely known error. Rater severity refers to the tendency for raters to consistently give higher or lower ratings than is justified by the performances (Engelhard, 1994). Differences in rater severity occur when raters do not interpret the rating scale in the same way, or have different standards or expectations. The same performance may be considered to be good, average, or poor by different raters

There are two other types of rater error that is central tendency and restriction of range. Central tendency happens when middle categories are used predominantly by raters. This judging behaviour often reflects the reluctance to use extreme categories. If ratings are somewhere in the average categories, there is a good chance that the ratings will not be too far from those given by another rater. Disagreement therefore becomes unlikely as the "implicit rule is when in doubt, avoid extreme categories" (Linacre, 1989). Cases of central tendency are typically detected by examining the pattern of category usage. Restriction of range, on the other hand, occurs when ratings are restricted to very few categories. Some raters may overuse the lower end of a scale while others may overuse the upper end. As restriction of range pertains to overuse of certain rating category, central tendency is, therefore, a special case of restriction of range. These two types of rater error are considered a serious threat to the quality of ratings as they fail to accurately discriminate examinees of different performance levels (Saal, Downey & Lahey, 1980).

A very severe or lenient rater may be considered to exhibit this kind or rater error. Another type of rater error relates to the internal consistency of ratings given by individual raters. Problems of internal consistency can be seen when raters are not consistent or constant in their judgment of similar performances. High ratings should be given to all good performances while low ratings should be given to all poor performances. Sometimes due to fatigue or inattentiveness, raters may award a high rating to a poor performance and a low rating to a good performance. Compared to rater severity, this type of rater error is considered to be more serious as raters are in themselves inconsistent in their judgment (Linacre, 1989).

Measurement situation becomes difficult situation when other aspects of the testing situation interpose themselves between the ability of the candidates and the difficulty of the test such as raters. Generally, there are two properties of judges' behavior that is leniency or severity of judges and the biasness. Therefore, the problem with intercorrelations between judge ratings is that they can just tell only the consistency among the rank of examinees' but do not inform about the severity or leniency differences between judges.

The popular measurement model used to deal with this performance assessment problem is the Many-facet Rasch Measurement Model. The objective is to develop basic measurement that can be used across same suitable measurement situations. Scores obtained via the Many-facet Rasch Model are believed to estimate accurately the students' ability. This is because Many-facet Rasch Model allows the use to separate students based on their abilities independently from other facets in the model such as tasks and raters (Engelhard &Myford, 2003). A student's raw score is adjusted for tasks difficulties and rater severity (Linacre, 1997)

Advances in theory and methodology are providing us with the framework and the tools to begin answering these questions and highlighting these problems. The notion of validity itself is being elaborated and extended. Newer research and statistical methods such as verbal protocol analysis (Ericsson & Simon, 1993) and item response theory (Hambleton, Swaminathan, & Rogers, 1991) are enabling us to find out what goes on in raters' minds and to tease out the different factors that affect rater ratings. This study uses one of these newer methodologies, the multi-facet extension of the Rasch model (Linacre, 1989), in conjunction with other research methods, to explore some of the challenges brought about by the use of performance assessments in the context of one particular exam, the Penilaian Kerja Amali(PEKA).

■3.0 PURPOSE OF STUDY

The Rasch Measurement Model is a powerful tool for handling polytomous data involving raters' judgements (Linarce, 1989). The values of separate facets, created on the same logit scale as person ability and item difficulty are estimated while the parameter separations is maintained. The Many-facet Rasch Model provides the simultaneous estimation of facet parameters so that they can be examined separately. Therefore, the purpose of this study is to examine how science teachers as the raters contribute to the scores of the primary schools students as the examinees in performing science process skill task as the items.

■4.0 METHODOLOGY

Research Design

This study uses the Many-facets Rasch Measuring Model to determine raters' severity and leniency. The research involves five raters and fifty examinees. It consists of standard five students of a primary school. There are five science process skills evaluated that is the observation, classifying, measuring, relationship between space and time and experimenting (identifying variables, make hypothesis, presenting the report). Three facets of the study-student ability, item difficulty and rater severity will be thoroughly discussed.

Research Participants

This research uses sample of 50 participants of standard five student comprising all males. The students consist of all races (25 Malay students, 3 Chinese students and 22 Indian students). Similar studies were done by Yang Ling Li (2004) whereby 5 raters and only 36 participants were involved. So, the usage of 50 participants justifies this research to enable the researcher to get a reliable result. Five raters (two males and three females) chosen are trained science teachers (it is either major or minor) in the school. All the raters were trained either by attending courses outside and went through 'in-house training' about rating the students in their performance assessment. They had taught the subject with minimum of at least three years experience. The raters will do the scoring after observing the pupils for the past two weeks (one experiment).

Research Procedures

The students were given one tasks for each session. Each sessions of experiment were conducted per week. The whole study took about 2 weeks to be completed. Each students need to complete the task in groups of four. Each student was given papers to answer the questions. These answer papers were given to the students as an attempt to get full participation from each student. After one hour, the experiment report were collected and given to the raters. Each rater rated each student twice using the scoring method given. 5 raters used the same timing to evaluate each student. This is to discard any outside factors that called contribute to any influence in their scoring. Each science process skill is rated in the holistic scoring rubric consists of a four-point rating scale where a score of one indicates the lowest level of performance and a score of four indicates the highest level of performance. The raters scored students' performance in the scoring forms given which consists of five Science Process Skills.

Data Analysis

The analyses were conducted using FACETS software. The design was a three-facet design consist students, tasks, and raters. The data from the rating scale was analyzed using Rasch Measurement Model (Wright & Stone, 1979) .Rasch analysis first converts the ordinal data generated by the rating scale into interval measure (Merbitz, Morris & Grip, 1989). Then, the analysis tests the interval validity of the scale by determining whether the items of the scale coalesce to form a single dominant construct or underlying dimension; this property is referred to as unidimensional (Haley, McHorney& Ware, 1994). Rasch analysis determines whether each participant was validly measured and whether each rater used the scale in a valid manner. The Rasch analysis also produces items calibrations that show how much of the underlying construct an item represents. Items with higher calibrations should be those, which are expected to represent more of the construct measured participants calibrations estimate the position of each participant assessed on the same continuum from less to more of the construct being measured and vice versa. Rater calibrations indicate how severe or lenient a rater is when assigning scores on the scale. A rater calibrated higher is more severe in assigning ratings and a rater calibrated lower is more lenient in assigning ratings. FACETS needed to be run more than once to answer the questions posed by this study. This FACETS analysis produced an overall scale expressed in terms of logits and all the facets were placed onto this scale, making meaningful comparisons between them possible. The FACETS measurement model allows the researcher to specify which of the various component factors in a measurement situation will have a predictable influence on the expression of ability; these factors are then removed from the final estimate of ability.

■5.0 RESULTS.

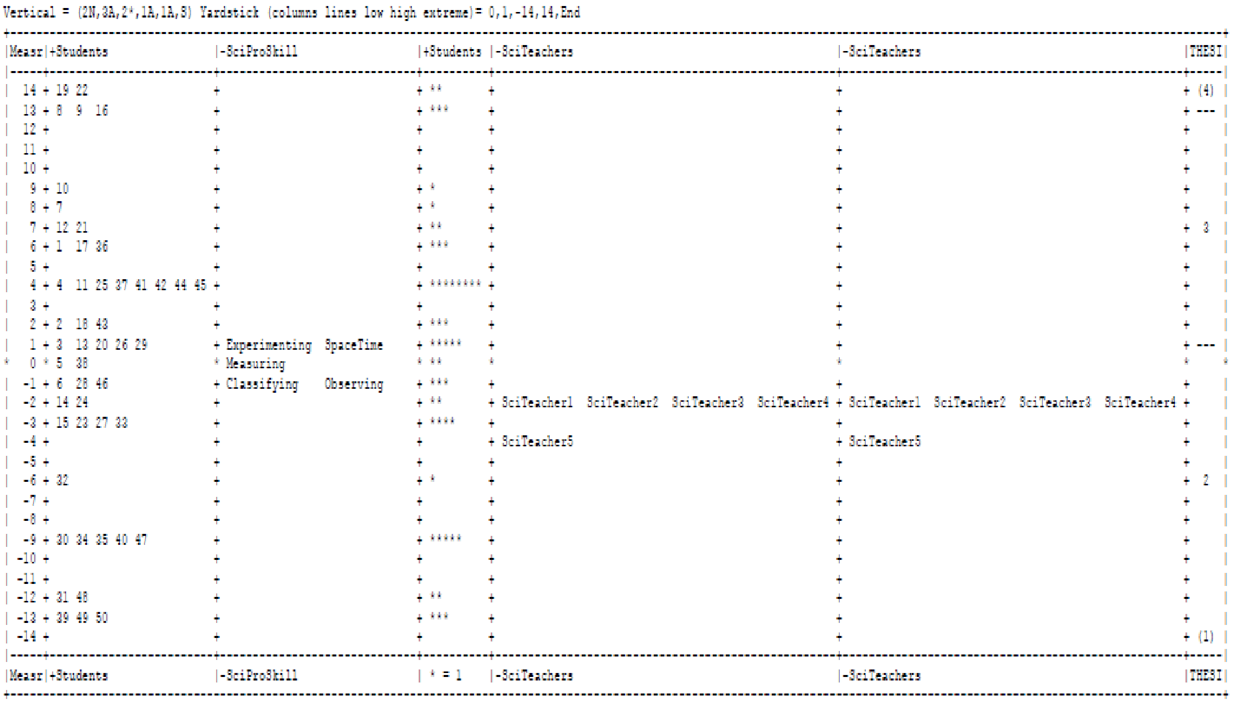
This purpose of the study was operationalized into two research questions dealing with raters, examinees and items and investigated through the use of multi-facet Rasch analysis. The data was analyzed with *Facets 3.68.0* a software program for MFRM(Linacre, 2011). Three facets were specified for this study: students, tasks, and raters types.

Findings For Research Question One : How Do The Science Teachers As Raters Differ In Terms Of Severity In Rating And Are Their Ratings Consistent?

Table 1 Data summary report

Assigning models to Data= "Peka.xls"
Total lines in data file = 254
Total data lines = 250
Responses matched to model: ?B,?B,?,PEKA,1 = 1250
Total non-blank responses found = 1250
Number of blank lines = 4
Valid responses used for estimation = 1250

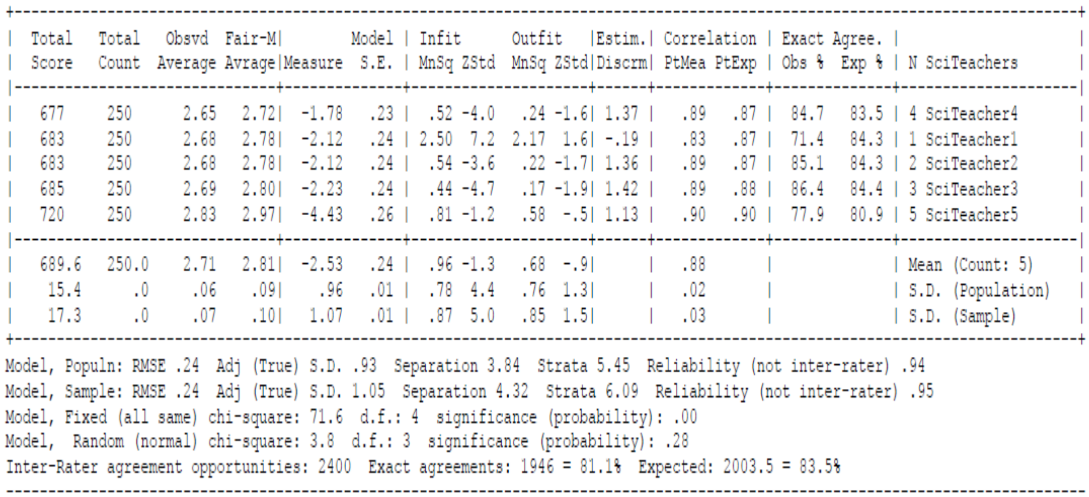
Table 2 All facets vertical ruler



```
Vertical = (2N,3A,2*,1A,1A,S) Yardstick (columns lines low high extreme)= 0,1,-14,14,End
+-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+Students                |-SciProSkill                |+Students |-SciTeachers                                        |-SciTeachers                                        |THESI|
|-----+-------------------------+----------------------------+----------+----------------------------------------------------+----------------------------------------------------+-----|
|  14 + 19 22                   +                            + **       +                                                    +                                                    + (4) |
|  13 + 8  9  16                +                            + ***      +                                                    +                                                    + --- |
|  12 +                         +                            +          +                                                    +                                                    +     |
|  11 +                         +                            +          +                                                    +                                                    +     |
|  10 +                         +                            +          +                                                    +                                                    +     |
|   9 + 10                      +                            + *        +                                                    +                                                    +     |
|   8 + 7                       +                            + *        +                                                    +                                                    +     |
|   7 + 12 21                   +                            + **       +                                                    +                                                    +  3  |
|   6 + 1  17 36                +                            + ***      +                                                    +                                                    +     |
|   5 +                         +                            +          +                                                    +                                                    +     |
|   4 + 4  11 25 37 41 42 44 45 +                            + ******** +                                                    +                                                    +     |
|   3 +                         +                            +          +                                                    +                                                    +     |
|   2 + 2  18 43                +                            + ***      +                                                    +                                                    +     |
|   1 + 3  13 20 26 29          + Experimenting  SpaceTime   + *****    +                                                    +                                                    + --- |
*   0 * 5  38                   * Measuring                  * **       *                                                    *                                                    *     *
|  -1 + 6  28 46                + Classifying    Observing   + ***      +                                                    +                                                    +     |
|  -2 + 14 24                   +                            + **       + SciTeacher1  SciTeacher2  SciTeacher3  SciTeacher4 + SciTeacher1  SciTeacher2  SciTeacher3  SciTeacher4 +     |
|  -3 + 15 23 27 33             +                            + ****     +                                                    +                                                    +     |
|  -4 +                         +                            +          + SciTeacher5                                        + SciTeacher5                                        +     |
|  -5 +                         +                            +          +                                                    +                                                    +     |
|  -6 + 32                      +                            + *        +                                                    +                                                    +  2  |
|  -7 +                         +                            +          +                                                    +                                                    +     |
|  -8 +                         +                            +          +                                                    +                                                    +     |
|  -9 + 30 34 35 40 47          +                            + *****    +                                                    +                                                    +     |
| -10 +                         +                            +          +                                                    +                                                    +     |
| -11 +                         +                            +          +                                                    +                                                    +     |
| -12 + 31 48                   +                            + **       +                                                    +                                                    +     |
| -13 + 39 49 50                +                            + ***      +                                                    +                                                    +     |
| -14 +                         +                            +          +                                                    +                                                    + (1) |
|-----+-------------------------+----------------------------+----------+----------------------------------------------------+----------------------------------------------------+-----|
|Measr|+Students                |-SciProSkill                | * = 1    |-SciTeachers                                        |-SciTeachers                                        |THESI|
+-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
```

Table 1 reports the numbers of observation which consists of 1250 responses (5 judges x 50 examinees x 5 traits =1250 responses). Table 2 shows the measures graphically. It can be observed that there is a noticeable spread among primary school students (examinees). However, there is a small spread among the science process skills performance(items) and science teachers(raters). The column heading the "-SciTeachers" informs that the most lenient judge will give the highest rating. "-means high score implies low measure", so "SciTeacher 5" is the most lenient rater.

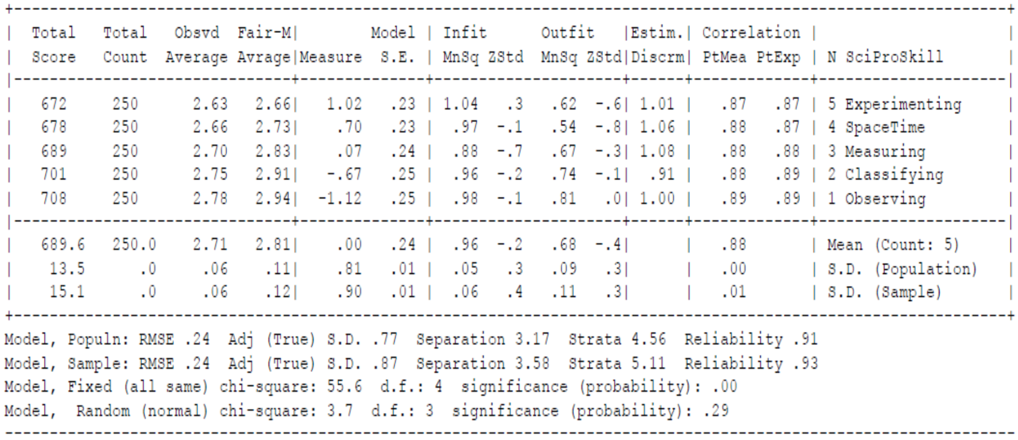
Table 3 Raters (science teachers) measurement report



```
+----------------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|        Model | Infit      Outfit    |Estim.| Correlation | Exact Agree. |                    |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | Obs %  Exp % | N SciTeachers      |
|-------------------------------+--------------+---------------------+------+-------------+--------------+--------------------|
|   677     250      2.65   2.72| -1.78    .23 |  .52 -4.0   .24 -1.6| 1.37 |   .89   .87 |  84.7   83.5 | 4 SciTeacher4      |
|   683     250      2.68   2.78| -2.12    .24 | 2.50  7.2  2.17  1.6| -.19 |   .83   .87 |  71.4   84.3 | 1 SciTeacher1      |
|   683     250      2.68   2.78| -2.12    .24 |  .54 -3.6   .22 -1.7| 1.36 |   .89   .87 |  85.1   84.3 | 2 SciTeacher2      |
|   685     250      2.69   2.80| -2.23    .24 |  .44 -4.7   .17 -1.9| 1.42 |   .89   .88 |  86.4   84.4 | 3 SciTeacher3      |
|   720     250      2.83   2.97| -4.43    .26 |  .81 -1.2   .58  -.5| 1.13 |   .90   .90 |  77.9   80.9 | 5 SciTeacher5      |
|-------------------------------+--------------+---------------------+------+-------------+--------------+--------------------|
|   689.6   250.0    2.71   2.81| -2.53    .24 |  .96 -1.3   .68  -.9|      |   .88       |              | Mean (Count: 5)    |
|    15.4      .0     .06    .09|   .96    .01 |  .78  4.4   .76  1.3|      |   .02       |              | S.D. (Population)  |
|    17.3      .0     .07    .10|  1.07    .01 |  .87  5.0   .85  1.5|      |   .03       |              | S.D. (Sample)      |
+----------------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .24  Adj (True) S.D. .93  Separation 3.84  Strata 5.45  Reliability (not inter-rater) .94
Model, Sample: RMSE .24  Adj (True) S.D. 1.05  Separation 4.32  Strata 6.09  Reliability (not inter-rater) .95
Model, Fixed (all same) chi-square: 71.6  d.f.: 4  significance (probability): .00
Model,  Random (normal) chi-square: 3.8  d.f.: 3  significance (probability): .28
Inter-Rater agreement opportunities: 2400  Exact agreements: 1946 = 81.1%  Expected: 2003.5 = 83.5%
------------------------------------------------------------------------------------------------------------------------
```

It is evident that there are differences in the raters' perception on the performance of primary school students in performing science process skills tasks. Raters severity varies about 2.65 logits (-4.43 to -1.78 logits) (Table 3). Science Teacher 4 relatively is the most severe rater (measure=-1.78 logits), whereas, the most lenient rater is Science Teacher 5 (measure=-4.43 logits). The statistical significance of judge variability is also examined by referring to the model, fixed (all same) chi square as suggested by Noor Lide Abu Kassim (2007). Thus, it is found that it is significant at p <0.01 (rater separation index of 4.32 and the chi square value of 71.6 with 4 degree of freedom). This gives an indication that judges consistently differ from one another in their overall severity of judgment. The observed number of exact agreement for all raters is 1946 (81.4%) out of 2400 rater agreement opportunities. This suggests that there is substantial agreement amongst judges as regards the placement of items by criterion points.

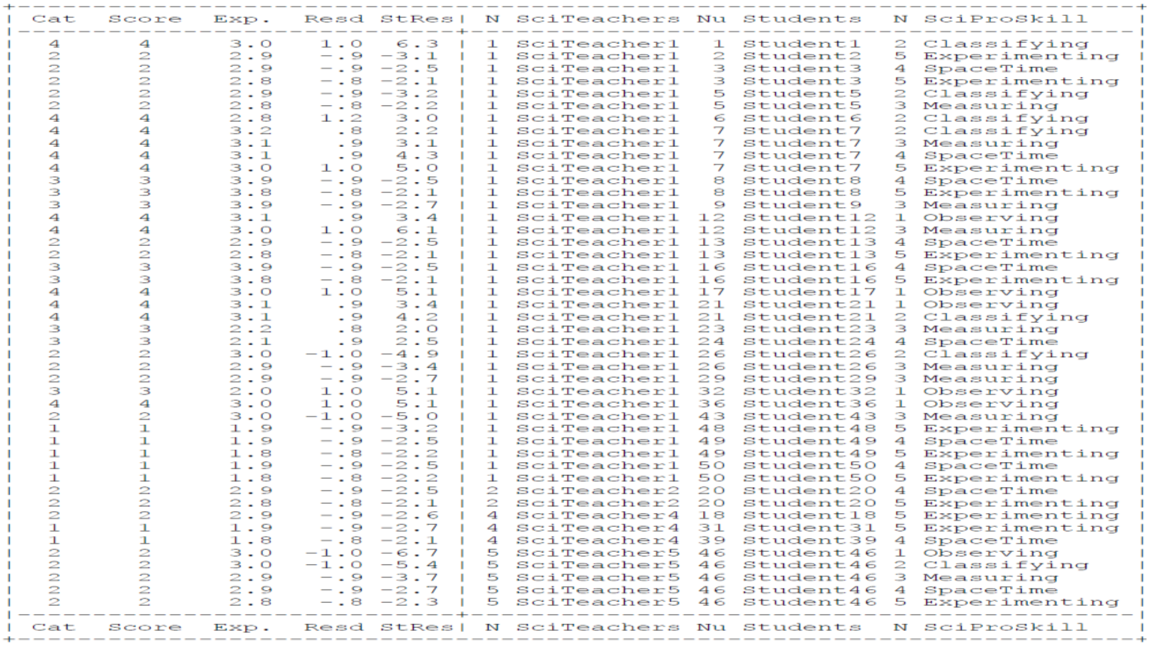
Table 4 Item (science process skill task) measurement report



```
+-----------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|        Model | Infit      Outfit   |Estim.| Correlation |                    |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | N SciProSkill      |
|-------------------------------+--------------+---------------------+------+-------------+--------------------|
|   672     250      2.63   2.66|   1.02   .23 | 1.04   .3   .62  -.6| 1.01 |   .87   .87 | 5 Experimenting    |
|   678     250      2.66   2.73|    .70   .23 |  .97  -.1   .54  -.8| 1.06 |   .88   .87 | 4 SpaceTime        |
|   689     250      2.70   2.83|    .07   .24 |  .88  -.7   .67  -.3| 1.08 |   .88   .88 | 3 Measuring        |
|   701     250      2.75   2.91|   -.67   .25 |  .96  -.2   .74  -.1|  .91 |   .88   .89 | 2 Classifying      |
|   708     250      2.78   2.94|  -1.12   .25 |  .98  -.1   .81   .0| 1.00 |   .89   .89 | 1 Observing        |
|-------------------------------+--------------+---------------------+------+-------------+--------------------|
|  689.6   250.0     2.71   2.81|    .00   .24 |  .96  -.2   .68  -.4|      |   .88       | Mean (Count: 5)    |
|   13.5      .0      .06    .11|    .81   .01 |  .05   .3   .09   .3|      |   .00       | S.D. (Population)  |
|   15.1      .0      .06    .12|    .90   .01 |  .06   .4   .11   .3|      |   .01       | S.D. (Sample)      |
+-----------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .24  Adj (True) S.D. .77  Separation 3.17  Strata 4.56  Reliability .91
Model, Sample: RMSE .24  Adj (True) S.D. .87  Separation 3.58  Strata 5.11  Reliability .93
Model, Fixed (all same) chi-square: 55.6  d.f.: 4  significance (probability): .00
Model,  Random (normal) chi-square: 3.7  d.f.: 3  significance (probability): .29
-------------------------------------------------------------------------------------------------------
```

PTMEA *Corr* value is used to determine an item polarity. If the PTMEA *Corr* is high, therefore the item posses the ability to differentiate the respondents ability (Table 4). Polarity Analysis for item towards a construct from the positive PTMEA *Corr* value shows the items in the construct is functioning in the same direction to measure the developed construct (Linacre, 2006). All the tasks in this study showed a highly positive PTMEA *Corr*. Experimenting receive the lowest average rating with the highest measure value of 1.02; meaning that its the most difficult science process task. Whereas, observing the highest rating with the lowest measure value of -1.12 . Therefore, observing is the easiest science process skill task.

Table 5 Unexpected responses



```
+------------------------------------------------------------------------------------------+
| Cat   Score   Exp.   Resd StRes| N SciTeachers Nu Students   N SciProSkill               |
|--------------------------------+---------------------------------------------------------|
|  4      4      3.0    1.0   6.3 | 1 SciTeacher1  1 Student1   2 Classifying              |
|  2      2      2.9    -.9  -3.1 | 1 SciTeacher1  2 Student2   5 Experimenting            |
|  2      2      2.9    -.9  -2.5 | 1 SciTeacher1  3 Student3   4 SpaceTime                |
|  2      2      2.8    -.8  -2.1 | 1 SciTeacher1  3 Student3   5 Experimenting            |
|  2      2      2.9    -.9  -3.2 | 1 SciTeacher1  5 Student5   2 Classifying               |
|  2      2      2.8    -.8  -2.2 | 1 SciTeacher1  5 Student5   3 Measuring                |
|  4      4      2.8    1.2   3.0 | 1 SciTeacher1  6 Student6   2 Classifying              |
|  4      4      3.2     .8   2.2 | 1 SciTeacher1  7 Student7   2 Classifying              |
|  4      4      3.1     .9   3.1 | 1 SciTeacher1  7 Student7   3 Measuring                |
|  4      4      3.1     .9   4.3 | 1 SciTeacher1  7 Student7   4 SpaceTime                |
|  4      4      3.0    1.0   5.0 | 1 SciTeacher1  7 Student7   5 Experimenting            |
|  3      3      3.9    -.9  -2.5 | 1 SciTeacher1  8 Student8   4 SpaceTime                |
|  3      3      3.8    -.8  -2.1 | 1 SciTeacher1  8 Student8   5 Experimenting            |
|  3      3      3.9    -.9  -2.7 | 1 SciTeacher1  9 Student9   3 Measuring                |
|  4      4      3.1     .9   3.4 | 1 SciTeacher1 12 Student12  1 Observing                |
|  4      4      3.0    1.0   6.1 | 1 SciTeacher1 12 Student12  3 Measuring                |
|  2      2      2.9    -.9  -2.5 | 1 SciTeacher1 13 Student13  4 SpaceTime                |
|  2      2      2.8    -.8  -2.1 | 1 SciTeacher1 13 Student13  5 Experimenting            |
|  3      3      3.9    -.9  -2.5 | 1 SciTeacher1 16 Student16  4 SpaceTime                |
|  3      3      3.8    -.8  -2.1 | 1 SciTeacher1 16 Student16  5 Experimenting            |
|  4      4      3.0    1.0   5.1 | 1 SciTeacher1 17 Student17  1 Observing                |
|  4      4      3.1     .9   3.4 | 1 SciTeacher1 21 Student21  1 Observing                |
|  4      4      3.1     .9   4.2 | 1 SciTeacher1 21 Student21  2 Classifying              |
|  3      3      2.2     .8   2.0 | 1 SciTeacher1 23 Student23  3 Measuring                |
|  3      3      2.1     .9   2.5 | 1 SciTeacher1 24 Student24  4 SpaceTime                |
|  2      2      3.0   -1.0  -4.9 | 1 SciTeacher1 26 Student26  2 Classifying              |
|  2      2      2.9    -.9  -3.4 | 1 SciTeacher1 26 Student26  3 Measuring                |
|  2      2      2.9    -.9  -2.7 | 1 SciTeacher1 29 Student29  3 Measuring                |
|  3      3      2.0    1.0   5.1 | 1 SciTeacher1 32 Student32  1 Observing                |
|  4      4      3.0    1.0   5.1 | 1 SciTeacher1 36 Student36  1 Observing                |
|  2      2      3.0   -1.0  -5.0 | 1 SciTeacher1 43 Student43  3 Measuring                |
|  1      1      1.9    -.9  -3.2 | 1 SciTeacher1 48 Student48  5 Experimenting            |
|  1      1      1.9    -.9  -2.5 | 1 SciTeacher1 49 Student49  4 SpaceTime                |
|  1      1      1.8    -.8  -2.2 | 1 SciTeacher1 49 Student49  5 Experimenting            |
|  1      1      1.9    -.9  -2.5 | 1 SciTeacher1 50 Student50  4 SpaceTime                |
|  1      1      1.8    -.8  -2.2 | 1 SciTeacher1 50 Student50  5 Experimenting            |
|  2      2      2.9    -.9  -2.5 | 2 SciTeacher2 20 Student20  4 SpaceTime                |
|  2      2      2.8    -.8  -2.1 | 2 SciTeacher2 20 Student20  5 Experimenting            |
|  2      2      2.9    -.9  -2.6 | 4 SciTeacher4 18 Student18  5 Experimenting            |
|  1      1      1.9    -.9  -2.7 | 4 SciTeacher4 31 Student31  5 Experimenting            |
|  1      1      1.8    -.8  -2.1 | 4 SciTeacher4 39 Student39  4 SpaceTime                |
|  2      2      3.0   -1.0  -6.7 | 5 SciTeacher5 46 Student46  1 Observing                |
|  2      2      3.0   -1.0  -5.4 | 5 SciTeacher5 46 Student46  2 Classifying              |
|  2      2      2.9    -.9  -3.7 | 5 SciTeacher5 46 Student46  3 Measuring                |
|  2      2      2.9    -.9  -2.7 | 5 SciTeacher5 46 Student46  4 SpaceTime                |
|  2      2      2.8    -.8  -2.3 | 5 SciTeacher5 46 Student46  5 Experimenting            |
|--------------------------------+---------------------------------------------------------|
| Cat   Score   Exp.   Resd StRes| N SciTeachers Nu Students   N SciProSkill               |
+------------------------------------------------------------------------------------------+
```

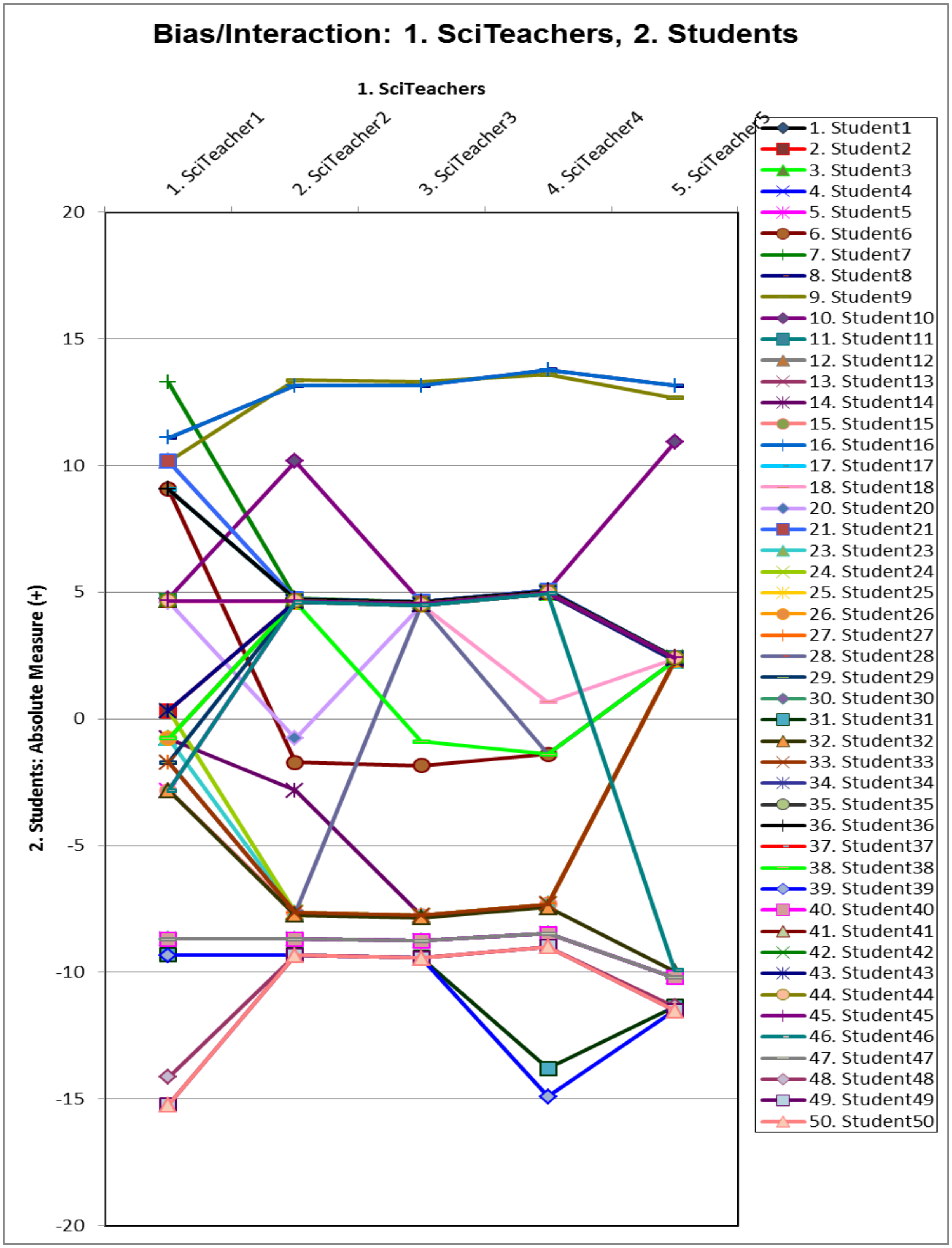


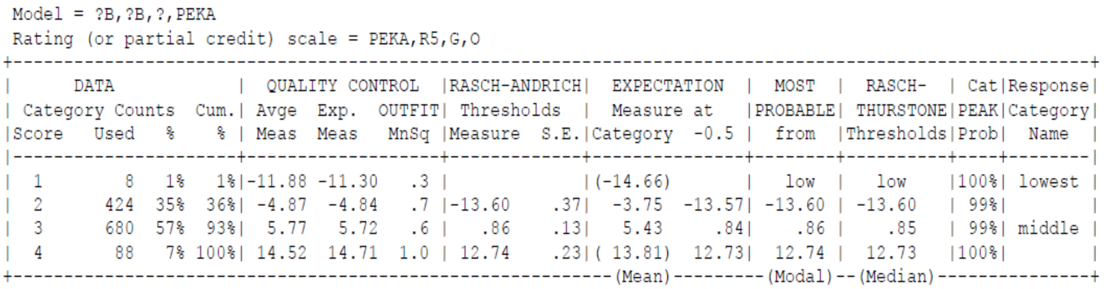
Figure 1 Bias/interaction analysis between science teachers (raters) and primary school students (examinees)

Science teacher 1 perception of the students is almost different from other raters. As an example, by referring the absolute measure, Science Teacher 1 give a very low rating to Student 5 compared to Science Teacher 2, 3,4 and 5 (Figure 1). On the other hand, Science Teacher 1 give a very high rating to Student 7 compared to Science Teacher 2, 3,4 and 5 (Figure 1). This supported by the results of unexpected response in Table 5 which informed that Science Teacher 1 has the highest unexpected responses compared to other science teachers in this study. Table 5 shows that Science Teacher 1 has 36 unexpected responses. As an example, Science Teacher 1 unexpectedly rated 4 to Student 1 when the student was performing the science process skill of classifying. Science Teacher 1 also unexpectedly rated 2 to Student 2 when the student was performing the science process skill of experimenting. On the other hand, there is only small number of unexpected responses made by Science Teacher 2, 4 and 5. Science Teacher 3 is the only rater which does not made any unexpected response.

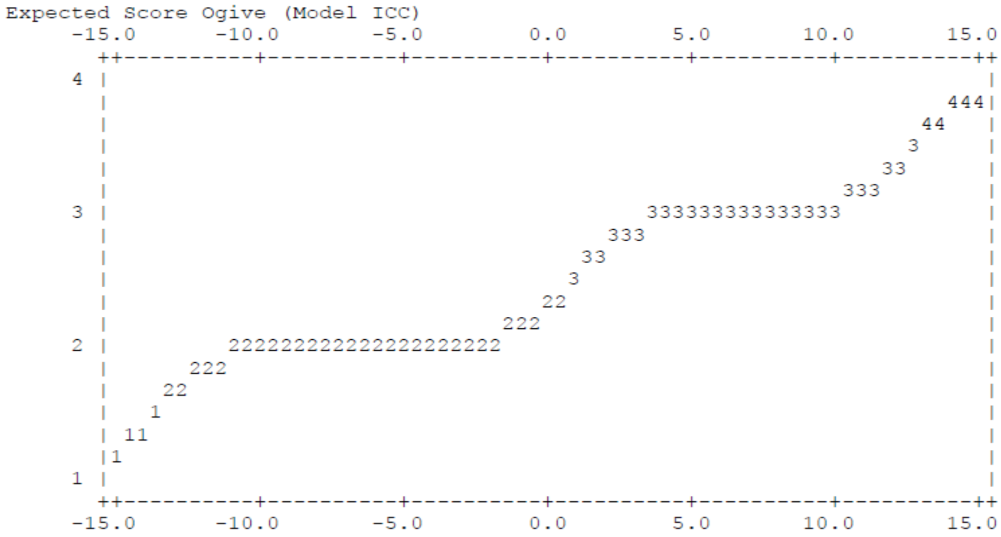
Findings For Research Question Two: How Is The Rating Scale Being Used By The Raters?

For each modeling rating scale, Facets served several pieces of information concerning to category use (Linacre, 2003b). The first information is related with the average measure of each category. Based on Table 6 and Figure 2, the results show that the average measures increase with each higher category which is aligned with the idea that the higher measures indicates the higher difficulty of the category (Linacre, 2003b).

Table 6 Science process skill measurement report



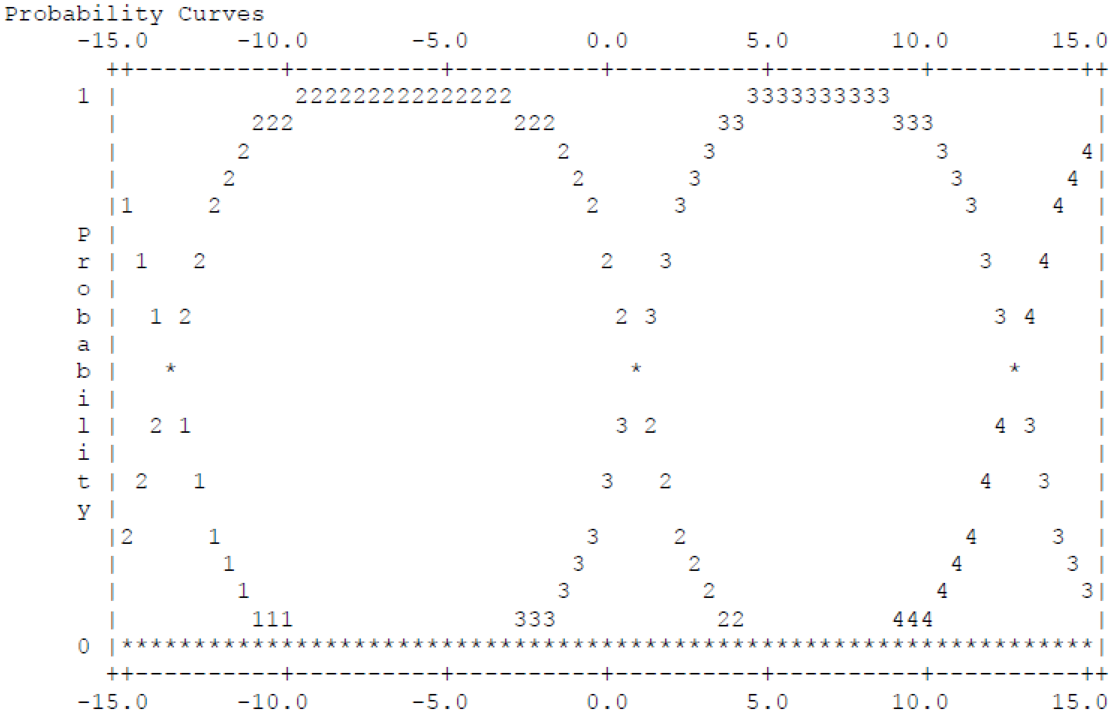
```
Model = ?B,?B,?,PEKA
Rating (or partial credit) scale = PEKA,R5,G,O
+-----------------------------------------------------------------------------------------------------+
|      DATA            |  QUALITY CONTROL  |RASCH-ANDRICH|  EXPECTATION  |  MOST  |  RASCH-   | Cat|Response|
| Category Counts  Cum.| Avge  Exp.  OUTFIT| Thresholds  |  Measure at   |PROBABLE| THURSTONE|PEAK|Category|
|Score   Used   %     % | Meas  Meas   MnSq |Measure  S.E.|Category  -0.5 |  from  |Thresholds|Prob|  Name  |
|----------------------+-------------------+-------------+---------------+--------+----------+----+--------|
|  1        8   1%   1%|-11.88 -11.30   .3 |             |(-14.66)       |   low  |   low    |100%| lowest |
|  2      424  35%  36%| -4.87  -4.84   .7 |-13.60    .37|  -3.75  -13.57| -13.60 | -13.60   | 99%|        |
|  3      680  57%  93%|  5.77   5.72   .6 |   .86    .13|   5.43     .84|    .86 |    .85   | 99%| middle |
|  4       88   7% 100%| 14.52  14.71  1.0 | 12.74    .23|( 13.81)  12.73|  12.74 |  12.73   |100%|        |
+-----------------------------------------------------------------(Mean)---------(Modal)--(Median)-----------+
```



```
Expected Score Ogive (Model ICC)
     -15.0       -10.0      -5.0        0.0        5.0       10.0       15.0
       ++----------+----------+----------+----------+----------+----------++
     4 |                                                                   |
       |                                                                444|
       |                                                             44    |
       |                                                            3      |
       |                                                          33       |
       |                                                       333         |
     3 |                                        333333333333333            |
       |                                     333                           |
       |                                   33                              |
       |                                  3                                |
       |                                22                                 |
       |                             222                                   |
     2 |         22222222222222222222                                      |
       |      222                                                          |
       |    22                                                             |
       |   1                                                               |
       | 11                                                                |
       |1                                                                  |
     1 |                                                                   |
       ++----------+----------+----------+----------+----------+----------++
     -15.0       -10.0      -5.0        0.0        5.0       10.0       15.0
```

Figure 2 Score ogive

The second information referring to observed use of each category. Two categories have noticeable higher frequency than the other categories (Table 6). The science teacher who ask to discriminate five levels in performing science process skill tasks, this data appear to be telling that a judge, could only discriminate two levels clearly which are score 2 and 3. The result is also supported by the Figure 3 which indicate the way in which the probability of scoring a lower category and the probability of scoring a higher category vary with ability (Wright and Masters, 1982). Probability curves that are prominent (clearly peaked) indicate clearly defined categories whereas probability curves that are less prominent indicate either narrowly defined categories or considerably improbable categories (Wright and Masters, 1982). Figure 3 showed that only score 2 and 3 are clearly peaked whereas score 1 and 4 is not clearly peaked. There is no peak at all for score 5. Thus, the result support that the science teachers as the rates can only discriminate two levels clearly which are score 2 and 3.



```
Probability Curves
     -15.0      -10.0       -5.0        0.0        5.0       10.0       15.0
       ++----------+----------+----------+----------+----------+----------++
     1 |            222222222222222                3333333333              |
       |         222               222           33          333           |
       |        2                     2         3               3         4|
       |       2                       2       3                 3       4 |
       |1     2                         2     3                   3     4  |
     P |                                                                   |
     r | 1   2                           2   3                     3   4   |
     o |                                                                   |
     b |  1 2                             2 3                       3 4    |
     a |                                                                   |
     b |   *                               *                         *     |
     i |                                                                   |
     l |  2 1                             3 2                       4 3    |
     i |                                                                   |
     t | 2   1                           3   2                     4   3   |
     y |                                                                   |
       |2     1                         3     2                   4     3  |
       |       1                       3       2                 4       3 |
       |        1                     3         2               4         3|
       |         111               333           22          444           |
     0 |*******************************************************************|
       ++----------+----------+----------+----------+----------+----------++
     -15.0      -10.0       -5.0        0.0        5.0       10.0       15.0
```

Figure 3 Probability curves

■6.0 DISCUSSION AND CONCLUSION

The possible threat to score validity identified by the study is the systematic variation typically built into science process skills assessments. In particular, different test takers respond to different science process skill tasks, and their responses are rated by different raters. The findings of this study suggest that scores can be affected if raters have different level of severity or leniency. Where science process skills are concerned, the results of this study suggest that in PEKA assessment, assigning science process skills to different test takers pose a threat to the validity of scores, which will make the test not valid and reliable.

Where raters are concerned, the results of this study suggest that raters of science process skills performance such as those in the PEKA can be trained to rate appropriately and consistently, and that under a system of double marking, assigning different raters to different test takers does not pose a threat to the validity of scores, and that tests are valid, reliable, and fair in that regard. The use of human raters also brings with it potential issues such as subjectivity and reliability, which could in turn affect the validity of test scores. The inter-rater reliability statistic only says something about the product of assessment but not about the process, and if we don‘t knows what raters are doing, then we don‘t know what their ratings mean (Connor-Linton, 1999). Raters could well be agreeing on things that have nothing to do with what is being measured. Thus, there is the need to better understand the rating process itself how raters go about the task of rating and what factors they actually consider as well as the rater characteristics that could affect raters' rating behavior.

As Lumley‘s (2005) model of the rating process shows, raters are an important component to the rating process. The rating process involves tension and struggle, as raters are people who come to the task of rating with different personalities and histories. And as these rater characteristics and backgrounds inform their rating, it is important to know what effects these have on the ratings they give.

Having considered the way raters of different experience and expertise rate, it is appropriate to consider whether their rating performance indeed differs.

As a science teacher who are assigned to evaluate the students performance in science performance assessment, are bound to evaluate the students' performances at one point or another. It could be an experiment, a portfolio, or a piece of writing. So, it is important to know that judging behavior can bring about unwanted variability or error in the measurement process and how these errors can affect the quality of ratings of students. Judging behavior must be examined and be conscious of how the raters rate the students' performances.

Eliminating rater errors cannot be done totally, but can be minimized in some ways. The use of a good measurement instrument is important. For example by using scoring rubric (analytical or holistic), the rating error can be reduce if characteristics to be rated represent specified learning outcomes, clearly defined and each rating scale describes the level of learning desired for an outcome. However, Ali Reza & Lovorn (2010) indicated that using rubrics may not improve the reliability or validity of assessment if raters are not well trained on how to design and employ them effectively. Rater training is another area that needed careful planning in order to minimized the rater error. An established body of literature shows that training can minimize rater effects. Latham, Wexley & Purcell (1975) used training to reduce rater effects among employment interviewers while Pulakos (1986) indicated that trained raters yielded more reliable (higher inter-rater agreement) and accurate (valid) ratings than no training.

Willingness to improve the ratings and to make the effort to ensure that the judgment of student performances is reliable and valid must be done cautiously and constantly. As assessment and its procedures are the core point of student learning (Lee King Siong, Hazita Azman, & Koo Yew Lie, 2010), matters related to valid and fair testing need to be taken seriously. It is hoped that with greater awareness of how the rating is done, teachers can be better raters and better teachers.

Although the findings of this study could be encouraging and promising, but certain limitations are gathered. This study was done quantitatively done. Lack of a qualitative component failed to provide us with convincing and justifiable reasons why the findings were obtained. The second limitation relates to the small number of teacher assessors in this research. Studies should be carried out in many schools and places in the country. There are co-ed schools, all-girls school and all-boys school. Other than that, there are many categories of schools such as vernacular, national, boarding schools. The location of the schools such as urban, semi-urban and rural schools do play a contributing factor that can influence the findings. Caution needs to be exercised as for the generalization of the results. The third limitation is the small number of items the raters rated (five items) and tested based on one topic. Future studies should include more items and covers more variety of topics in the science syllabus of primary schools. The duration of the assessment should be done in a longer period .This is to justify the finding. Every examinee is given enough time to proof themselves. The last limitation concerns about students language proficiency. Instructions and wordings used in the assessment should be made simple so that the language does not become another variable that could distort the findings. Finally, due to the small sample size we cannot really make any statements about whether rater assessment could be a reliable. Future studies should strive to answer this important question.

Performance assessment is regarded as subjective type of assessment compared to other type assessment because human judgment is involved in giving the scoring for examinees. Sometimes, the human factor involved in scoring may cause the scores not to be totally trustworthy. So, issues such as reliability and validity of scores did by raters are main concern when dealing with performance assessment. So, it is important to know that judging behavior can bring about unwanted variability or error in the measurement process and how these errors can affect the quality of ratings of students. Judging behavior must be examined and be conscious of how the raters rate the students' performances.

References

Bond, T. G. & Fox, C. M, (2007). *Applying The Rasch Model: Fundamental Measurement In The Human Sciences*. Second Edition. Ney Jersey: Lawrence Erlbaum Associates Publishers.
Cambridge, MA: MIT Press

Connor-Linton, J. (1999). Competing communicative styles and crosstalk: A multi-feature analysis. *Language in Society*, 28(01), 25-56.
Dahncke, R. Duit, W. Gräber, M. Komorek & A. Kross, Eds, *Research in Science Education – Past, Present And Future*, 49-60. Dordrecht: The Netherlands: Kluwer Academic Publishers.

Engelhard, G. (1994). Examining rater errors in the assessment if written composition with a many-faceted Rasch Model. *Journal of Educational Measurement*, 31(2), 93-112.

Engelhard, Jr&Myford, M.C (2003).*Monitoring Faculty Consultant Performance In The Advanced Placement English Literature And Composition Program With A Many-Faceted Rasch Model*. (College Board Research Rep. No. 2003-1). New York.

Englehard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, 5(3), 171-191.

Englehard, G. (1994). Examining rater errors in the assessment of written compositions with a many-faceted Rasch model. *Journal of Educational Measurement*, 31(2), 93-112.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. 2nd edition.

Grobman, Laurie. 2007. Affirming The Independent Researcher Model:Undergraduate Research in the Humanities. *CUR Quarterly,* 28(1*)*:23-28

Haley, S.M, McHorney, C.A, &Ware,J.E., Jr (1994), Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. *Journal of Clinical Epidemiology*, 47, 671-684.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals Of Item Response Theory*. Newbury Park, CA: Sage.

Harlen, W. (1999). The assessment of scientific literacy in the OECD/PISA project. In H. Behrendt, H.

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18(2), 5-17.

Lee King Siong,HazitaAzman& Koo Yew Lie. (2010). Investigating the undergraduate experience of assessment in higher education. *GEMA Online Journal of Language Studies*, 10(1), 17-33.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (1997). *Guidelines for rating scales. MESA Research Note #2.* Retrieved June 24, 2009, from http://www.rasch.org/rn2.htm

Linacre, J. M. (2002). What do infit, outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16, 878.

Linacre, J. M. (2006). *Facets Rasch measurement computer program*. Chicago: Winsteps.com.

Linacre, J. M. (2011). *Winsteps Rasch Measurement Version 3.71* [Software].

Linacre, J.M. (2003) *Winsteps Version 3.4 Wright & Master* (1982). Rating Scale Analysis. Chicago: MESA

Linacre, J.M. (2003). *Winsteps Version 3.48* [Computer Software and manual]. Chicago.

Lumley T (2005). *Assessing Second Language Writing: The Rater's Perspective.* Frankfurt.

McNamara, T. F. (1996). *Measuring Second Language Performance*. London: Longman.

McNamara, T.F. (1996). *Measuring Second Language Performance*, New York; Longman.

Merbitz, C., Morris, J., & Grip, J. (1989). Ordinal Scales and the foundations of misinference. *Archieves of Physical Medicine and Rehabilitation*, 70, 308-332.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3).

Moon, T., & Calahan, C. (2001). Classroom performance assesment: What should it look like in a standards-based clasroom? *NASP Buletin*, 85(62), 48-58.

Noor Lide Abu Kassim(2007). *Using The Rasch Measurement Model For Standard Setting Of The English Language Placement Test At The IIUM.*

Pulakos, E.D. (1986). The Development Of Training Programs To Increase Accuracy On Different Rating Forms. *Organizational Behavior And Human Decision Processes*, 38, 76-91.

Rasch, G. (1980). *Probabilistic Models For Some Intelligence And Attainment Tests*. Chicago: University of Chicago Press.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18-39.

Skamp, K.(1998). Primary Science & Technology: How Confident Are Teachers? Research in Science Education, 21, 290-299.
Test Method And Learner Discourse. *Language Testing*, 16(1), 82-111.

Upshur, J.A & Turner CE (1999). *Systematic Effects In The Rating Of Second Language Speaking Ability:*

Wiggins, G (1993). *Educative Assessment.* Designing Assessment to inform and improve student Performance, San Francisco, California.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, 13, 181-208

Wright, B. & Stone, M. (1979). *Best Test Design*. Chicago: MESA

Wright, B. D., & Masters, G. N. (1982). *The measurement of knowledge and attitude*. Research Memorandum No. 30. Chicago: University of Chicago, MESA Psychometric Laboratory.

Yang Ling Li (2004), *Psychometric Properties of the Volitional Questionnaire*, University of Illinois, Chicago.