RCEE & RHEd2010
Kuching,Sarawak
7 – 9 June 2010

The Measurement of Teaching Effectiveness in Engineering Education using Rasch Analysis

Omar, **M.Z** [a], **Rodzo'an**, **N. A**[b], **Saidfudin**, **M** [c] , **Zaharim, A** [d] and **Basri, H** [e]

[a]Assoc.Prof., Centre for Engineering Education Research,
Fac. of Engineering and Built Environment, Universiti Kebangsaan Malaysia, 43600 Bangi, MALAYSIA
[b] Master (*Candidate*), Dept. of Mathematics, Fac. of Science,
[c] Program Director, Exec.Dip. in Quality Management, UTM SPACE,
University Teknologi Malaysia, 81300 Skudai, MALAYSIA
[d]Professor, Centre for Engineering Education Research, Fac. of Engineering and Built Environment
[e] Professor, Ir; Deputy Vice Chancellor (A) ,Universiti Kebangsaan Malaysia, 43600 Bangi, MALAYSIA

Abstract

The Engineering Accreditation Council of Malaysia (EAC) adopts the American Accreditation Board of Engineering and Technology 2000 (ABET) requirements which promote outcome based education (OBE) learning process. OBE calls for the evaluation of the subjects learning outcomes (LO) as specified in the Programme Specification. This good practice is implemented in the Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia (UKM) teaching and learning processes. Evaluation method has been largely dependent on students' performance carrying out tasks such as tests, quizzes or submission of assignments. Instrument construct were based on Bloom's Taxonomy whilst the evaluation on the students performance output were assessed based on Students Observed Learning Outcomes (SOLO) taxonomy which gives an indication on the student achievement of the subject expected LO. However, the measurement of the students achievement from the observed outcomes remain vague. This paper describes a measurement model using Rasch Analysis which can be used to measure the subject LO of an undergraduate engineering subject. An overview of the measurement model and its key concepts are presented and illustrated using the final exam paper given through subject KKKF1134 – Introduction to Engineering. The final examination results were evaluated on how well it relates to the latent trait abilities being scrutinised whether it correspond to the LO that is to be measured. Attributes for each dimension were duly identified and responses were coded polytomously to clearly define the assessment rubrik. Results obtained were assessed against the course LO maps for consistency and used as a guide for future improvement of the teaching method and style. The study shows that Rasch model of measurement can classify grades into learning outcomes more accurately especially in dealing with small number of sampling unit.

Keywords: Learning Outcomes, instructional objectives, performance assessment, Quality, continuous improvement.

1. Introduction

The assessment of student learning begins with educational values. Assessment is not an end in itself but a vehicle for educational improvement. Its effective practice, then, begins with and enacts a vision of the kinds of learning we most value for students and strive to help them achieve. Educational values should drive not only *what* we choose to assess but also *how* we do so. Where questions about educational mission and values are skipped over, assessment threatens to be an exercise in measuring what's easy rather than a process of improving what we really care about [1].

Assessment is most likely to lead to improvement when it is part of a larger set of conditions that promote change. Assessment alone changes little. It's greatest contribution comes on campuses where the quality of teaching and learning is visibly valued and worked at. On such campuses, the push to improve educational performance is a visible and primary goal of leadership; improving the quality of undergraduate education is central to the institution's planning, budgeting, and personnel decisions. On such campuses, measuring learning outcomes to generate useful and meaningful information about is seen as an integral part of decision making, and avidly sought. [2]

The Engineering Accreditation Council of Malaysia (EAC) adopts the American Accreditation Board of Engineering and Technology 2000

(ABET) requirements which promote outcome based education (OBE) learning process. OBE calls for the assessment of the subjects learning outcomes (LO) as specified in the programme specification. IHLs in Malaysia conducting any engineering programmes must assess the learning outcomes of its teaching and learning processes as a prerequisite to obtain EAC accreditation hence measurement.

2. Overview of Measurement Principles

Measurement has been grossly misunderstood and overlooked in many circumstances especially in the field of social science. Many researchers in social science are frustrated when existing instruments are not well tailored to the task, since they then cannot expect sensitive, accurate, or valid findings [3]. However, modern measurement method as practiced using item response theory with a focus on Rasch measurement model provides the social sciences with the kind of measurement that characterizes measurement in the natural sciences; i.e. the field of metrology.

The fundamentals of measurement must comprised of the instrument to be used for purpose which has specific unit of an agreed standard amount. An instrument must have the correct construct of linear scale which can be zero set and duly calibrated. A valid instrument can then be replicated for use independent of the subject hence measurement taken thereof is therefore a reliable data for meaningful analysis and examination to generate useful information [4]. This information is of utmost importance to be the prime ingredient in a particular decision making.

3. Measurement Method

Responses from the students in an examination, test or quizzes is normally marked against a marking scheme comprising keywords; where when there is a match then the student would be given a mark or otherwise. This is the traditional 'park and mark system'. In theory, at this stage truly the assessors is only counting the number of correct answers which is then added up to give a total raw score. The raw score is only give a ranking order which is deemed an ordinal scale that is continuum in nature. It is not linear and do not have equal intervals which contradicts the nature of data fit for the due statistical analysis [5]. It does not meet the fundamentals of sufficient statistics for evaluation.

In Traditional Test Theory, these data set would normally be put on a scatter plot to establish the best regression. However, estimate or prediction from ordinal responses on the student learning outcomes (LO) attributes are almost impossible due to the absence of equal interval on a linear scale. The normal solution in linear regression approach is to establish a line which fits the points as best as possible; which is then used to make the required predictions by inter-polation or extra-polation as necessary as shown in Figure 1.

$$y = \beta_0 + \beta_1 m \qquad \text{Equ. (1)}$$

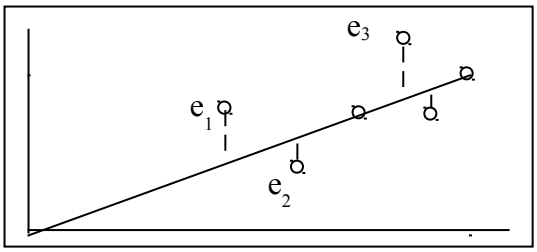


Figure 1 –Best fit line concept

In obtaining the best fit line; however, there exist differences between the actual point; y_i, and the predicted point; $\acute{y}_i$, that is on the best fit line,. The difference is referred here as error; e

$$y_i - \acute{y}_i = e_i \qquad \text{Equ. (2)}$$

By accepting the fact that there is always errors involve in the prediction model, the deterministic model of equation (1) renders itself less reliable. This can be resolved by transforming it into a probabilistic model by including the prediction error into the equation;

$$y = \beta_0 + \beta_1 m + e \qquad \text{Equ. (3)}$$

Rasch moves the concept of reliability from establishing "best fit line" of the data into producing a reliable repeatable measurement instrument [14] instead. Rasch focuses on constructing the measurement instrument with accuracy rather than fitting the data to suit a measurement model with of errors. By focusing on the reproducibility of the latent trait measurement instead of forcing the expected generation of the same raw score, i.e. the common expectation on repeatability of results being a reliable test, the concept of reliability takes its rightful place in supporting validity rather than being in contentions. Hence; measuring LO ability in an appropriate way is vital to ensure valid quality information can be generated for meaningful use; by absorbing the error and representing a more accurate prediction based on a probabilistic model.

In Rasch philosophy, the data have to comply with the principles, or in other words the data have to fit the model. In Rasch point of view, there is no need to describe the data. What is required is to test whether the data allow for measurement on a linear interval scale specifically in a cumulative response process i.e. a positive response to an item stochastically implies a positive response to all items being easy or otherwise. This is dichotomous

responses which can take only two values, 0 and 1 which is known as Bernoulli random variable; in our case a smart student or otherwise.

Rasch Measurement Model is expressed as the ratio of an event being successful as;

$$P(\theta) = \frac{e^{(\beta n - \delta i)}}{1 + e^{(\beta n - \delta i)}} \qquad \text{Equ.(4)}$$

where;

e = base of natural logarithm or Euler's number; 2.7183
βn = person's ability
δi = item or task difficulty

Rasch exponential expression is a function of Logistic Regression which resulted in a ***S***igmoidal ogive and can be transformed into simpler operation by reducing the indices by logarithm:

$$\ln[P(\theta)] = \ln\left[\frac{e^{(\beta n - \delta i)}}{1 + e^{(\beta n - \delta i)}}\right] \text{ Eq (5)}$$

Now ln[P(θ)]; as the probability of a successful event; x=1 is reduced to the expression in equation 6 and can be construed simply as the difference of person ability; β_n and the item difficulty; δ_I, which can be represented as;

$$\ln [P(\theta)] = \beta_n - \delta_i ; \qquad \text{Equ.(6)}$$

The very reason why the need to transformed it to *logit* is primarily to obtain a linear interval scale. It can be readily shown mathematically that a series of numbers irrespective of based used is not equally spaced but distant apart exponentially as the number gets bigger while a log series maintain their equal separation; thus equal interval [6]. This equal separation is mathematically shown in Table 1. The difference between $\log_{10}5$ and $\log_{10}2$ is constant and remain of equal distant between $\log_{10}50$ and $\log_{10}20$ which similarly hold true for $\log_e$; thus the theorem can now be universally applied.

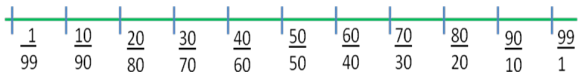
Table 1. Comparison of Numerical and Log intervals

Numerical series	$\log_{10}$	$\log_e$
1	0.000	0.000
2	0.301	0.694
5	0.699	1.609
10	1.000	2.303
20	1.302	2.997
50	1.699	3.912
100	2.000	4.606

Similarly, an attempt of a student to answer a question can be seen as a chance of him being able to get the correct answer or successfully accomplishing a given task. Now, for a given normal score of 7/10 which is normally read as 70%; there is need of a paradigm shift to read it as the odds of success being 70:30; thus a ratio data. A mark of 6/10 shall now be seen as odd of success 60:40 and, so on. After all percentage is statistically recognized only a data summary; which is somehow largely confused as a unit of measurement [7].

This enable us to construct a log-odd ruler of probability an event taking place with the odd-of success as shown in Figure. 2 with unit termed as ***logit*** , derived from the term '**log-odd unit**'; as unit of measurement of ability akin to *meter* to measure length or *kilogram* to weight.

Figure 2. Probabilistic line diagram



In order to achieve an equal interval scale, we can introduce logarithm of the odd probabilistic value. Maintaining the same odd probabilistic ruler as in Figure 2, starting with 0.01 to 100, we can create an equal interval separation between the log odds units on the line, hence the measurement ruler with the *logit* unit [8]. This can be verified by computing the value of $\log_{10}$ 0.01 (10^{-2}) equals to -2.0; value of $\log_{10}$ 0.1 equals to -1; value of $\log_{10}$ 1 equals to 0 and so forth. Figure 3 shows the newly established *logit* ruler as a linear scale with equal interval separation. It is just like looking at a thermometer with '0', as water being ice and 100 as boiling point whilst the negative extreme end as -273°C, the point where all atoms of any element come to a standstill.

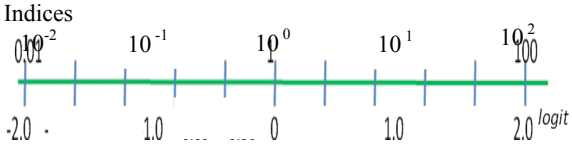


Figure 3. Logit ruler

Thus, we now have a valid construct of an instrument to measure the students ability for each defined LO.

4. Results and Discussion

The test was administered on 1st year Engineering and Architecture students from the Faculty of Engineering and Built Environment, University Kebangsaan Malaysia (UKM) for the course code KKKF1134 – *Introduction to Engineering*. The result from the tests were tabulated and ran in *Winsteps v 3.6.8,* a Rasch

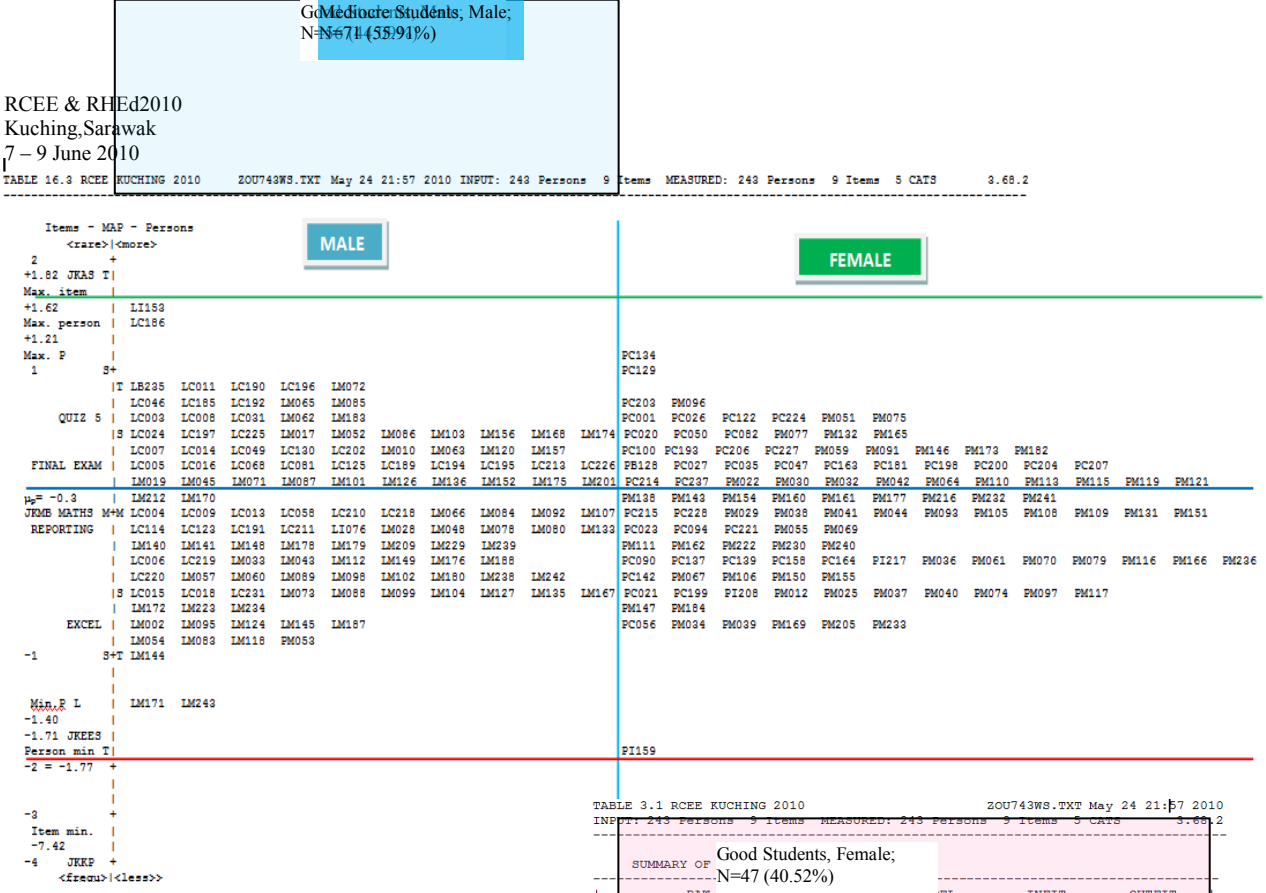


Figure 4 -Person-Item Distribution Map

based analysis software; to obtain the *logit* values. Figure 4 shows the Person-Item Distribution Map (PIDM) where the *persons*; i.e. the Students is on the right whilst the *items*; the learning topics were plotted on the left of the *logit* ruler as in Figure 3. By virtue of the same ruler with the same scale; then the correlation of the *person*, β_n and *item*, δ_i can now be established as in equation (6).

The PIDM Map is the heart of Rasch analysis [9]. The vertical dashed line represents the ideal less-to-best continuum of quality. Items and students now share the same linear measurement units known as logits. On the left hand side of the dashed line, the items are aligned from too easy to too hard, starting from the bottom. The distribution of student positions is on the right side of the vertical dashed line in increasing order of ability; the best naturally being at the top and the poorest student is at the bottom of the rung. Letter "M" denotes the student and item mean, "S" is one standard deviation away from the mean and "T" marks two standard deviations away from the mean. In Rasch Model, since we are interested in the person's ability for a given task, it is most prudent to zero set the scale where the item mean is zero when the ability is deemed 50:50 being the tipping point.

Figure 2 shows the PIDM: *Students Location* where the students were separated by gender to evaluate their trend in learning. Rasch Analysis tabulates the students' location in a very clear graphical presentation which is easy to read and easier to understand. Each student can be coded with attributes or factors that is deemed to affect their learning process.

Students Location

This will enable in depth analysis of their study pattern to be evaluated meaningfully. Before delving any further, it is best to look at the analysis Summary Statistics as in Table 2. The prime information we are looking for in this table is the overall students' LO ability reflected by the Person Measure ; μ_{PERSON}= -0.03*logit* (P[Θ]=0.4925). This gives the indication that generally the students performance under scrutiny is just slightly below expectation. SD=0.48 shows that the students is very much within target though we noted that poor students are LM171, LM242 with PI159 measured -1.77*logit* being lowest whilst the best students are PC134, PC129, LC186 and LI152 measured at +1.62*logit* being topmost .

Table 2. Summary Statistics

Both student's ability measurements can give us some indication where are the students are on the probability scale; -ve means they are to the left of the 'thermometer', +ve means they are located to the right of the scale. Now we can sense and have a better appreciation if the students are in trouble or not since now their performance is duly measured on sound metrology principles thus generasibility.

There are 127 Males against 116 Females students in this study. Male students shows a mean of, μ_M=-0.02*logit* where they were found to performed slightly better than their Female counterparts with a lower mean of, μ_F=-0.04*logit*. Generally, the students separation, G=0.66 is such a small value that indicates that there is not enough differentiation among students ability to separate them into distinct performance level or strata. Strata can be calculated using the formula:

Strata = (4 X student separation +1)/3 Equ. 7

Thus, a student separation of 0.66 was computed into the strata formula which yielded 1.21; only a poor 2 separated groups (good, mediocre). This is clearly reflected by the PIDM in Figure 4. Generally, it shows the students into two (2) separate profiles;

Group 1: Mediocre students; (Male,N=71, 55.91%; Female,N=69,59.48%) who has complication attempting the Final Exam, Quiz 5 and JKAS. They encounter some troubles pursuing Mathematics Fundamentals, JKMB and preparing the Reports.

Group 2: Good students; (Male, N=56, 44.09%; Female, N=47,40.52%) who has good command of all subject matter but some difficulty in attempting Quiz 5 and complexity in JKAS.

Further scrutiny of the students responses shows lack of partial knowledge amongst the students. Table 3 shows the pattern of responses. The structure calibration; '*s'* is assessed to confirm the rating classification used is true where *s*-value being the separation between each structure category label;

e.g; $s_{2\text{-}3}$= -0.88-0.27 =-0.61; < 1.4, ***Not*** OK.

$s_{3\text{-}4}$= 0.57-(-0.88) =1.45; < 1.4, OK.

The separation shall be in the range where s; 1.4< *s*<5. It is noted in Figure 5 that the difference for each category are irregular where the difference between category 2, 3 and 4, 5 are all less than 1.4. It can be seen that classification 2 is well submerged and 4 just below 3and 5. Therefore, the ability classification A,5>90; B, 4>80; C,3>70; D,2>60 and Fail,1<60 is not reflective of this cohort person separation. Hence, we need to re-classify the rating and, in Rasch this is termed as collapsing. On the other hand if s> 5, then the category cluster need to be split instead. Subsequent to the re-scoring, if it is found that the SD is larger, then the new re-scoring will taken as the better measurement. Otherwise, it shall be retained.

This pattern of dichotomous response gave rise to concern of students lacking partial knowledge. In engineering, prudent engineers must possess some partial knowledge particularly in design concept or engineering philosophies. A mechanical mindset alone does not suffice and they need a more rigorous program to change their mindset to be 'ingenious'.

Table 3. Structure Calibration

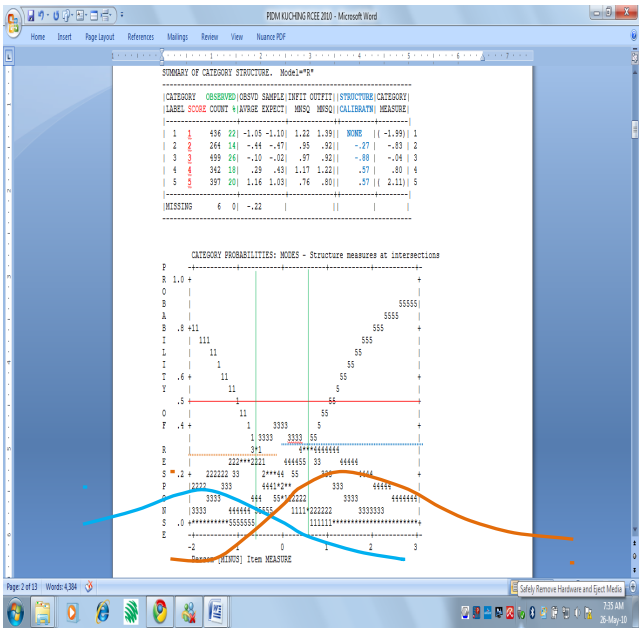
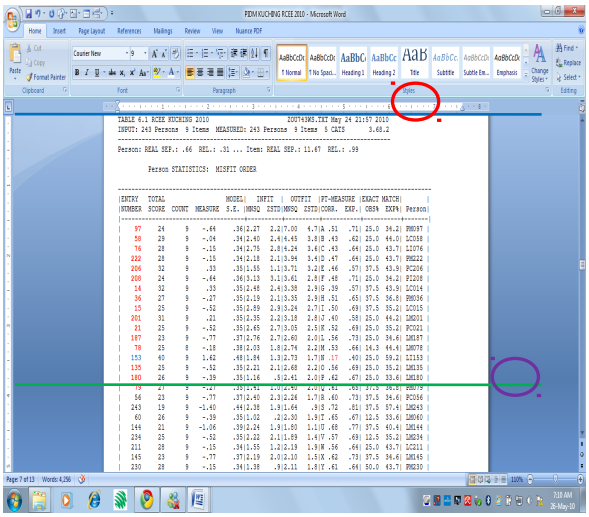


Figure 5. Item Characteristic Curve

Rasch has a unique ability in recognizing the students development based on the students responses. Table 4 shows the Person Measure Fit Order. This table gives an indication the validity of the person responses whether it fits the model; i.e. the fundamental of Rasch Model .

Table 4. Person Fit Order



Student PM097 attempted all the 9 domains but obtained only 24 out of possible 45 since each domain has a maximum rating 5. Coded

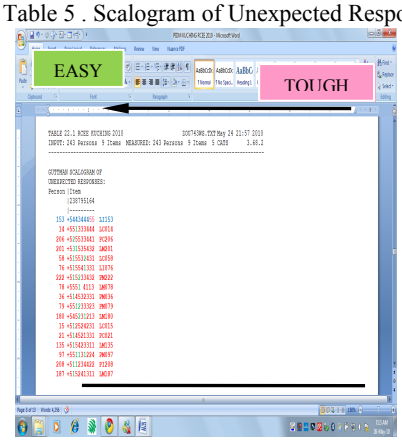
demographically as PM; "Perempuan Melayu", her ability is measured as -0.64 *logit* which means she is on the lower side of the scale.

Rasch examine item or person fit by looking at two types of fit values known as infit and outfit Rasch typically examine 'outfit' which is less threatening to measurement and easier to manage. Hence, we look at "outfit MNSQ" where the mean square (MNSQ) outfit for the students is expected to be near 1.0. Acceptable MNSQ outfit shall be between 0.5 and 1.5.

Table 4 gives the Person misfit responses; the topmost being worst where the data provided are outfit to the model thus multi-dimensionality. PM097 shows MNSQ Outfit=7.00 which means a far item is not correctly assessed. Closer examination of the scalogram pattern response in Table 5 shows that item 8, 7 and 5 are somehow not well assessed and need to be reviewed. Those rating 1 could have been more appropriately be higher value. Similarly, item 4 should have been a 1 instead of 4. Rasch would ask the researcher to identify the reasoned argument 'why' does this happen.

However, the point measure correlation can give more interesting pattern of responses. Though the acceptable value is in the range of 0.38 to 0.85, perhaps we shall start to worry as it approaches near zero or make it worst negative value. It simply means the respondent is behaving the opposite way. Let us have a look at Table 5 Student LI153. He appeared to have scored a high 40 out of 45 from 9 items, but with a high MNSQ Outfit= +2.73 and a very low point measure correlation=+0.17, which is near zero. Analysing his pattern of response in Table 5 Scalogram confirms that he is peculiar. Take note he faired poorly on the easy items, on the left but scored well on the difficult items on the right. Despite his high score, Rasch identified his pattern of responses is different that warrants justification thus response validity.

Table 5 . Scalogram of Unexpected Responses



These outcomes does not meet Rasch model expected outcomes. This major finding raised some conclusions, for example, the student underestimated the easiest items hence careless errors. Conversely, for the difficult items, suspects probably have special interest or knowledge on the topic and/or comfort answering statement-based question. On the other hand, it makes sense that the student may simply guess the answers for the questions. Rasch has this particular predictive properties embedded in the model to make it a very reliable validation model.

5. Conclusion

Rasch Model provides a sound platform of measurement equivalent to natural science which matches the SI Unit measurement criteria where it behaves as an instrument of measurement with a defined unit and therefore replicable. It is also quantifiable since it's linear. Rasch Model has made it very useful with its predictive feature to overcome missing data [3].

The logit ruler has been developed with purpose to measure ability; in this case students learning ability of specific learning outcomes. It can define the students profile and most important we are now able to validate a question construct on line. It is a noble innovation where the ability 'ruler' can transform ordinal data into measurable scale. It's graphical output is great which gives better clarity for quick and easy decision making.

The measurement conducted reveals the true degree of cognitive learning abilities of the Engineering undergraduates based on Blooms Taxonomy [10]. Previously, lack of such measurement in Engineering Education has made the necessary corrective actions in the form of skills development, education and competency training difficult to formulate. This major problem faced by Engineering Education Administrators in an IHL to design the necessary curriculum to mitigate the going concern is therefore resolved. Rasch has all the capabilities to rigorously analyse examination results more accurately thus making evaluation clearer to read and easier to understand [11].

Acknowledgements

The authors wish to acknowledge the financial support received from the Centre for Engineering Education Research, University Kebangsaan Malaysia as research grant in the effort of improving the quality of teaching and learning in engineering education.

References

[1] B. D. Wright and M. M. C. Mok, "An overview of the family of rasch measurement models," in *Introduction to Rasch Measurement: Theory, Models, and Applications*, J. Everett V.Smith and R. M.Smith, Eds., 2004, p. 979

[2] Astin. A.W, et. al. 9 Principles of Good Practice for Assessing Student Learning, The American Association for Higher Education, 1991, Stylus Publishing, LLC, 2005

[3] Saidfudin, M., and Azrilah, A.A., , "Structure of Modern Measurement", Rasch Model workbook Guide, ILQAM, UiTM, Shah Alam. 2009. Retrievable at http://www.ilqam.uitm.edu.my/wp-content/uploads/2009/06/0.1-Rasch-Workshop-Booklet_Structure-of-Measurement.doc

[4] Saidfudin, M. and Ghulman, H . A; "Modern measurement paradigm in Engineering Education: Easier to read and better analysis using Rasch-based approach", *International Conference on Engineering Education*, ICEED2009, Dec. 9-10, Shah Alam

[5] Sick, J. Rasch Measurement in Language Education Part 3: The family of Rasch Models, Shiken: *JALT Testing & Evaluation SIG Newsletter* Vol. 13 No. 1 Jan. 2009 (p. 4 - 10) [ISSN 1881-5537]

[6] Saidfudin, M, Azlinah M , Azrilah AA, Nor Habibah, A. & Sohaimi Z, "Appraisal of Course Learning Outcomes using Rasch measurement: A case study in Information Technology Education", *International Journal of Systems Applications, Engineering & Development*; Issue 4, vol.1, University Press, UK. pp.164-172, July 2007

[7] Saidfudin, M, Rozeha, A , Razimah A. & Hamza A Ghulman, " Application of Rasch-based ESPEGS Model in Measuring Generic Skills of Engineering Students: A New Paradigm", in *WSEAS Transactions on Advances in Engineering Educatio*n, Issue 8 Vol.5, WSEAS Press. pp. 591-602, August 2008.

[8] Saidfudin, M, Azlinah M , Azrilah AA, NorHabibah, A; Hamza A Ghulman & Sohaimi Z, "Application of Rasch Model in validating the construct of measurement instrument", in *International Journal of Education and Information Technologies,* Issue 2, Volume 2,. pp. 105-112; May 2008

[9] Saidfudin, M., Azrilah, A.A., Azlinah, M.,Nor Habibah, A., Zakaria, S., and H.A. Ghulman, "Development of Rasch-based Descriptive Scale in profiling Information Professionals' Competency", in *IEEE XPLORE* indexed in INSPEC; *IEEE IT Simposium* (ITSim KL),2008, pp.329-333, August 2008.

[10] A.Chapman, "Bloom's Taxonomy - Learning Domains." vol. 2007: Businessballs.com, 2006

[11] Saidfudin, M. Invited Paper, "Intelligent Students' Learning Ability Measurement System: Easier to understand and clearer analysis using Rasch Model", *Proceedings of 2009 IEEE International Conference on Antennas, Propagation and Systems* (INAS 2009), 3-5 Dec. 2009, Johor, Malaysia