Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 60 (2012) 228 – 235

Procedia
Social and Behavioral Sciences

UKM Teaching and Learning Congress 2011

Aligning assessment with learning outcomes

Nangkula Utaberta[a,b]*, Badiossadat Hassanpour[a,b]

[a]*Department of Architecture, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia*
[b]*Centre for Engineering Education Research, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia*

Abstract

Architecture is a multidisciplinary field of study that draws on the arts, science and social sciences. However, the most important part of architectural education in terms of curriculum focus and time spent by students is architectural design. It is in the design studio that students are expected to bring together knowledge from the different disciplines to inform the development of their architectural designs. The design studio offers the potential to provide a multifaceted and enriching learning experience. The crit or project review is a form of teaching and assessment to which schools of architecture have subscribed for decades and this historical continuity would seem to suggest that in the past it has been a successful mode of transmitting the knowledge and skills of the architect to the next generation of the profession. But continuity of a social institution may reflect more than functional effectiveness. Using the inherited models and method of evaluation and assessment in design studios is the crucial part of architecture design studios. This paper tries to align the assessment model with learning outcomes trough out Rasch measurement model.

Keywords: Assessment; architecture Studio; learning outcomes

1. Introduction

The notion of culture represents a fundamental basic human concept that underlies historical developments and the creation of civilizations. Though there are numerous ways of defining culture, it is often perceived as referring to the shared ways of thinking and behaving, to common attitudes and beliefs that a social community shares , and to the products the social community has created. The concept of culture is currently applied broadly to refer to, depict and characterize sets of shared beliefs and modes of practice in diverse areas, including in the sphere of education (For instance learning cultures, school culture) and educational assessment.

The role of assessment is shifting. Assessment currently perceives as a means to promote learning rather than monitor it, hence assessment is for learning. Assessment for learning is the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there.

* Corresponding author. Tel.: +6-03-8921-7035; fax: +6-03-8921-6147
E-mail address: nangkula_arch@yahoo.com

1877-0428 © 2011 Published by Elsevier Ltd. Selection and/or peer reviewed under responsibility of the UKM Teaching and Learning Congress 2011
doi:10.1016/j.sbspro.2012.09.372

Assessment should not merely seen as something separable from instruction, administered at the end of the learning process, but also as a powerful tool for promoting deep learning activities. Each of the assessments is important – those that are occur in daily classroom interactions among teachers and students, those set by teachers at the end of particular phase in the work, and those developed and administered by external jurors. Assessment culture refers to educational evaluation practices that are compatible with current ideologies, social expectations, attitudes and values so the emergence of assessment cultures needs to be discussed with reference to current views on learning and education and the social role of assessment. Birenbaum 1996 has made a distinction between cultures in the measurement of achievement and relates them to the developments in learning society. In the traditional so-called testing culture, instruction and testing are considered to be separate activities. Instruction is the responsibility of the teacher, whereas testing is the responsibility of the psychometric expert, who can use elaborate procedures for test development and sophisticated psychometric models for the analysis of test responses. The changing learning society has generated the so-called assessment culture as an alternative to the testing culture.

Learning cultures approaches are rooted in Piaget's cognitive development theory and in situated practice theories, which perceive learning as "an integral and inseparable aspect of social practice. The changes are taking place with increasing moves towards what we call Powerful Learning Environments (PLEs). A PLE is characterized by a good balance between discovery learning and personal exploration on the one hand, and systematic instruction and guidance, on the other. It also takes into account the students' individual differences in abilities, needs, and motivation. The assessment culture strongly emphasizes the integration of instruction and assessment. Students play far more active roles in the evaluation of their achievement. The construction of tasks, the development of criteria for the evaluation of performance, and the scoring of the performance may be shared or negotiated among teachers and students. The assessment takes all kinds of forms such as: observations, text- and curriculum-embedded questions and tests, interviews, performance assessments, writing samples, exhibitions, portfolio assessment, and project and product assessments. From a psychometric point of view, they can be characterized as badly standardized assessments that depend heavily on instruction. In these new contexts, assessment will be used more frequently, and will have different functions, different goals, and different administration techniques in addition to new assessment forms.

Communities which endorse the assumptions of learning cultures recognize intelligence as multi-faceted and aim to provide opportunities for all students to learn in modes consistent with their linguistic, idiosyncratic and social background, without prior labeling of predicted capacities. According to Birenbaurn 1996, the assessment culture is in accord with the constructivist approach to education. In this approach, learning is viewed as a process through which the learner creates meaning and teacher is not a person who transfers knowledge, but a mentor who provides opportunities for learners to use the knowledge and skills they already possess in order to understand new topics. The teacher is expected to provide interesting and challenging tasks. Teymur in 1985 reached to this conclusion that the design process consists of regular experimentation, it can be said that architectural curriculum generally has few real variations in different countries .but the reason which leads to distinction in results are the differences in learning style which originate from numerous cultures and variety of backgrounds. The research consisted of 108 university first-year Bachelor's students as sample by David Gijbel in 2006 shows that differences in assessment preferences are correlated with differences in approach to learning. Deep approaches to learning are associated with students' intentions to understand and construct the meaning of the learned content, whereas surface approaches to learning refer to students' intentions to learn by memorizing and reproducing the factual contents of the study materials. Students generally shift between surface and deep approaches to suit the assessment demands of their courses. In this paper the effectiveness of formative assessment would be discussed in architecture education the by using Rasch measurement model and second year design studio at Universiti Kebangsaan Malaysia tried to evaluate the current assessment model and find the weak and strength point to use as a lever for improvement.

2. Effect of Formative Assessment on Education

Assessment of students' achievement is an important factor in encouraging students to adopt these kinds of deep learning strategies. There are several studies that have demonstrated the positive relationship between a deep approach to learning and study success. The well-known study by Marton and Saljo (1976) gave qualitative evidence for this, but later studies also provided more quantitative evidence. Deep learners prefer courses which are intellectually challenging and assessment procedures allowing them to demonstrate their understanding.

The conditions for good practice are having manageable workload; regular feedback; and structural support to provide an overview of contents in order to distinguish the wood from the trees and not going beyond the capabilities of the students. The topic of formative assessment has been extensively reviewed by Black and William 1998. According to these authors, all those activities undertaken by teachers and/or by their students that provide information to be used as feedback, in order to modify the teaching and learning activities in which they are engaged, can be labeled as modes of formative assessment. Increasing formative assessment will lead to the following effects:

1. Reactivating or consolidating prerequisite skills or knowledge prior to introducing new material
2. Focusing attention on important aspects of the subject
3. Encouraging active learning strategies
4. Giving students opportunities to practice skills and to consolidate learning
5. Providing knowledge of outcomes and corrective feedback
6. Helping students to monitor their own progress and to develop self-evaluation skills
7. Guiding the choice of further learning activities to increase performance
8. Helping students to feel a sense of accomplishment

Nitko identifies four types of instructional decisions that are supported by tests:

1. Placement decisions:

Deciding in which instructional sequence or at what level in the instructional sequence a student should begin in order to avoid unnecessary repetition of what is already known and to allow more rapid attainment of new goals. Placement or prior knowledge state tests are given before a student begins a unit of instruction. They may focus on the prerequisite knowledge and skills for proposed learning or on the outcome knowledge and skills to determine whether students have already attained the desired outcome of the instruction. Thus, instructors use prior knowledge state tests to place students in instructional modes by (a) determining the degree to which prerequisite entry behaviors or skills are present or absent, (b) determining entering mastery of course objectives, and also for (c) matching students to alternative instructional modes, based on student characteristics.

2. Monitoring decisions:

Deciding (a) whether students are learning appropriately, and (b) whether the assigned learning activity is working effectively, or a different activity should be assigned. So-called progress assessments or assessments of growth are the instruments for the assessment of a student's progress toward the final learning objectives. Progress assessments provide feedback to students on how their learning is progressing and can lead to the (formative) decision that there is a need to remediate or relearn a certain instructional unit. Progress assessments are used by instructors to (a) choose or modify subsequent learning activities, (b) prescribe remediation of group or individual deficiencies, and (c) provide on-going feedback to the student for the purpose of directing advanced or remedial study. Progress tests are taken during instruction, but measure student behaviors and cognitive processes at the level of the desired outcomes.

3. Attainment decisions:

Deciding, at the end of a particular instructional segment, whether a student has attained the desired instructional goals. Final tests are used for certifying student learning, and evaluating teacher effectiveness.

4. Diagnostic decisions:

Deciding which learning outcomes a student has not acquired and the probable cause(s) of the failure to acquire them, in order to remediate or correct incomplete or erroneous prior learning. Diagnostic assessments are designed to provide specific information on individual learning deficiencies and misunderstandings. When prior knowledge state assessments, progress assessments, and final assessments are designed and interpreted properly they can also provide information of a diagnostic nature since students and teachers can use this information in order to regulate learning processes. Nisbet 1993 defines the term authentic assessment as methods of assessment which influence teaching and learning positively in ways which contribute to realizing educational objectives, requiring realistic (or

authentic) tasks to be performed and focusing on relevant content and skill. Essentially similar to the tasks involved in the regular learning processes in the classroom.

3. Case study and Research method

According to Hussein, academic excellence is student's achievements which are based on university's assessments such as test, assignment, presentation, final exam and etc. Assessment should reflect these understandings by employing a diverse array of methods, including those that call for actual performance, using them over time so as to reveal change, growth, and increasing degrees of integration. Such an approach aims for a more complete and accurate picture of learning, and therefore firmer bases for improving our students' educational experience. If students are to be prepared for practice and develop life-long learning skills, then assessment that depends on traditional review or crit should be re-examined. The review should be more closely reflect the range of skills needed by architects in professional practice, with particular reference to communication with clients and users. It should develop and build these skills cumulatively during the undergraduate course. Providing a clearer basis by defining more explicitly than before, the required learning outcomes, the weight of assessment and criteria for assessment can be helpful for improvement.

To monitor and evaluate the quality and efficiency of proposed model in chapter three, which is based on criteria-based assessments' rules and qualifications, final submission of the third project which was designing a medium size building in a dense urban context has chosen as case study the reason is that, the chosen project is the most comprehensive assignment with maximum and clear objectives and defined tasks. 23 students were presented their designed projects to the jurors. The students were using Auto cad and Archi Cad to present their design. Before the submission day students have been informed about the objectives of the projects and assessment tasks which were going to be assessed by jurors. Also an evaluation sheet has been prepared for all the jurors that were containing the objectives and criteria for the assessment. The defined evaluation sheet for this submission day was included 3 main criteria which were included oral and graphic presentation, design development, and model.
Each of these has defined into different tasks for marking. The tasks are as below:

- Oral and graphic presentation:
  - Attire and composition
  - How clear is the information
  - Focus and explanation

- Design development:
  - Study on architectural language and understanding of the issue
  - Understanding of the precedent study
  - Concept and idea development
  - Design approaches
  - Respond to the site
  - Spaces and spatial integration
  - Completeness of drawings (sketches and technical drawing)
  - Building proportion, scale, texture, colour, and how it is composed

- Model:
  - Completeness of the model
  - Detail of the model, proportion, scale, texture and colour
  - The use of materials, finishing and detailing

The importance of each objective and task also has defined by percentage and level of satisfaction of jurors in each task was defined from fail, poor, average, good to excellent. Four jurors were attending in submission day , one of them were PhD holder with more than 10 years experience (teacher 2), the other one was a senior lecturer with more than 15 years experience in practice and academics (teacher 4), and two young lecturers with 3years (Teacher 1) and 1 year experience (Teacher 3) respectively. Before starting the jury session, they had a meeting with

master of the studio and discussed about the project and objectives in detail and each of the jurors received 23 evaluation form. Each student has given 10 to 15 minutes to explain his idea and the development process and planning details. After that jurors got time to ask questions and give comments. Finally the evaluation forms were collected from all jurors to be the base of total mark to students.

4. Rasch measurement model and mini Facet Software

The development of Rasch Measurement Model in social science educational measurement has rapidly expanded to other areas of education including technical and engineering fields. And the problem can be solved with use of Rasch measurement model in architecture too. Rasch moves the concept of reliability from establishing 'best fit line' of the data into producing reliable repeatable measurement instrument. This measurement model uses empirical data directly from the lecturer's assessment on student for a given task and transformed them into logic scale which have equal interval. Rasch analysis can be applied to assessments in a wide range of disciplines, including health studies, education, psychology, marketing, economics and social sciences. Rasch models are used for analyzing data from assessments to measure variables such as abilities, attitudes, and personality traits. For example, they may be used to estimate a student's reading ability from answers to questions on a reading assessment. Rasch models are particularly used in psychometrics, the field concerned with the theory and technique of psychological and educational measurement. Analyzing data according to the Rasch model, that is, conducting a Rasch analysis, gives a range of details for checking whether or not adding the scores is justified in the data. This is called the test of fit between the data and the model. To evaluate the data from the studio we used Mini Facet software. The key in data has tabulated base on each student with different jury. So 1302 digit have had key in to the software.

5. Findings and Recommendations

To key in the data first of all we turned the names of criteria to shorten form and contract to name them with first alphabet of each task and numbers. Defined codes are as mentioned below:

I. Oral and graphic presentation as (O) group: The tasks named like and Attire and composition (O1), how clear is the information (O2), focus and explanation (O3).

II. Design development as (D) group: The tasks named Study on architectural language and understanding of the issue (D1), Understanding of the precedent study (D2), Concept and idea development (D3), Design approaches (D4), Respond to the site (D5), Spaces and spatial integration (D6), Completeness of drawings (sketches and technical drawing) (D7), Building proportion, scale, texture, colour, and how it is composed (D8).

III. Model as (M) group: Completeness of the model (M1), Detail of the model, proportion, scale, texture and colour (M2), the use of materials, finishing and detailing (M3). Also for jurors we named the participants as teacher 1 for the juror who was phD candidate with 3 years experience. Teacher 2 for PhD holder with more than 10 years experience, Teacher 3 for the youngest teacher with one year experience and master degree, and teacher 4 for a senior lecturer with more than 15 years experience in practice and academics. Students also turned from s1 to s23. The result from the final submission were tabulated and run in *Mini facets*, Rasch analysis software; to obtain the logit values. Figure 1 shows the summary statistics of students' measurement. Under this figure is a set of reliability statistics. These show the reliability of the differences between the measures in the Facet. They indicate the reproducibility of the measures, not the accuracy of the measures. These reliabilities are not inter-rater reliability statistics (which show the rater similarity). Reproducible means we can expect the same number if we repeated the same data collection. A stopped clock is highly reproducible, so it is highly reliable. Of course, it is reliably wrong! Model means assuming all misfits in the data is due to the randomness predicted by the Rasch model. Real means assuming all misfits in the data contradicts the Rasch model. Population means assuming this set of elements is the entire population. Sample means assuming this set of elements is a random sample from the population of interest. RMSE means root mean-square error, a statistical average of the standard errors of the measures. Adj (True) SD means the standard deviation of the measures, (Adj=) adjusted for measurement error, also called the "True" standard deviation. Separation is the True SD / RMSE. It

indicates how many measurement strata could be statistically distinguishable among the measures, if the tails of the measure distribution are conceptualized to be caused by outlying random noise.

Reliability is the ratio of the "True" variance of the measures to the observed variance. Strata is (4*Separation + 1)/3. It indicates how many measurement strata could be statistically distinguishable among the measures, if the tails of the measure distribution are conceptualized to be caused by outlying "true" measures. As figure 1 shows student reliability is at 0.97 which is above 0.7 indicating that the student sample has a good spread of students' ability, in which the students can be separated into about 6 separations. The most performed student is located at 2.85 logit student16. The least performed student is located at -3.08logit that is student20. It means that six group of students are allocated in the good rating scale. Fit statistics by Rasch Analysis enable the researchers to see whether the data they are using is feasible or not. Specifically in persons' capability and item difficulty, since it is desirable to at the person fit as well as the item fit, an irregular or erratic responses could be a sign of misfit. A misfit item means that the particular item is either too difficult or too easy for respondent or it could mean that the item is not really testing on the desired latent trait. There are means of checking for quality control in Rasch. In order to verify for fit or misfit items, the following criteria must be satisfied: Point Measurement Correlation: $0.32 < x < 0.8$ - Outfit Mean Square: $0.5 < y < 1.5$ - Outfit Z Standard: $-2.0 < Z < +2.0$

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu	STUDENTS
101	28	3.6	3.44	2.85	.36	.43	-2.3	.42	-2.4	1.50	.71	.58	16	student16
139	42	3.3	3.34	2.52	.28	1.44	1.8	1.40	1.6	.61	.66	.67	15	student15
98	28	3.5	3.33	2.47	.35	1.55	1.8	1.52	1.7	.44	-.18	.57	8	student8
97	28	3.5	3.29	2.35	.35	1.99	3.1	1.97	3.0	-.11	.02	.57	12	student12
136	42	3.2	3.27	2.29	.28	.64	-1.7	.58	-2.1	1.44	.83	.67	5	student5
143	42	3.4	3.25	2.22	.28	1.62	2.7	1.56	2.4	.45	.79	.56	1	student1
90	28	3.2	3.07	1.54	.34	.78	-.9	.78	-.9	1.33	.22	.58	6	student6
83	28	3.0	2.85	.76	.34	.54	-1.9	.55	-1.8	1.41	.69	.61	13	student13
122	42	2.9	2.02	.64	.20	1.13	.6	1.12	.5	.90	.64	.62	7	student7
40	14	2.9	2.81	.61	.48	.24	-2.6	.27	-2.2	1.60	.84	.65	2	student2
108	42	2.6	2.68	.21	.27	.89	-.4	.84	-.6	1.14	.58	.70	9	student9
77	28	2.8	2.63	.07	.34	.96	.0	.88	-.2	1.12	.87	.66	14	student14
75	28	2.7	2.54	-.16	.34	1.16	.6	1.08	.3	.93	.21	.66	11	student11
94	42	2.2	2.29	-.77	.26	.93	-.2	.90	-.3	1.04	.74	.70	3	student3
101	42	2.4	2.23	-.92	.26	1.08	.4	1.04	.2	.83	.67	.66	10	student10
64	28	2.3	2.04	-1.34	.32	.56	-1.9	.55	-1.9	1.54	.86	.66	4	student4
63	28	2.3	2.00	-1.44	.32	.64	-1.5	.64	-1.5	1.37	.85	.65	17	student17
28	14	2.0	1.62	-2.31	.43	.51	-1.6	1.00	.1	.98	.00	.61	22	student22
48	28	1.7	1.49	-2.66	.31	.85	-.5	.93	-.1	1.13	.50	.51	21	student21
48	28	1.7	1.42	-2.88	.31	.96	.0	.83	-.4	1.14	.55	.51	19	student19
47	28	1.7	1.39	-2.98	.32	1.35	1.3	1.11	.4	.36	.50	.50	18	student18
46	28	1.6	1.36	-3.08	.32	.28	-4.0	.26	-2.8	1.86	.75	.49	20	student20
84.0	31.2	2.7	2.51	.00	.32	.93	-.4	.92	-.3		.56		Mean (Count: 22)	
32.9	8.4	.6	.70	1.96	.05	.44	1.8	.42	1.6		.30		S.D. (Population)	
33.6	8.6	.6	.72	2.01	.05	.45	1.8	.43	1.6		.31		S.D. (Sample)	

Model, Popuin: RMSE .33 Adj (True) S.D. 1.93 **Separation 5.88** trata 8.17 **Reliability .97**
Model, Sample: RMSE .33 Adj (True) S.D. 1.98 Separation 6.02 Strata 8.36 Reliability .97
Model, Fixed (all same) chi-square: 822.2 d.f.: 21 significance (probability): .00
Model, Random (normal) chi-square: 20.5 d.f.: 20 significance (probability): .43

Figure 1. Student's Measurement Report

Figure 2 shows criteria measurement report. Item reliability is >0.7 given as 0.96 to indicate that the items have a good spread of item difficulties to measure the design progress of students' project.

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu	CRITERIA
93	49	1.9	1.58	2.43	.26	1.17	.8	1.02	.1	.94	.77	.76	13	M2 most difficult
93	49	1.9	1.58	2.43	.26	1.30	1.3	1.14	.5	.80	.77	.76	14	M3
97	49	2.0	1.68	2.16	.26	1.62	2.5	1.55	1.9	.45	.74	.77	12	M1
137	49	2.8	2.71	-.31	.25	1.04	.2	1.05	.2	.85	.63	.75	1	O1
137	49	2.8	2.71	-.31	.25	.53	-2.8	.53	-2.8	1.49	.85	.75	6	D3
137	49	2.8	2.71	-.31	.25	.70	-1.6	.73	-1.4	1.27	.83	.75	7	D4
139	49	2.8	2.75	-.44	.25	.78	-1.1	.82	-.9	1.17	.78	.75	8	D5
140	49	2.9	2.77	-.50	.25	.80	-1.0	.76	-1.2	1.22	.74	.75	3	O3
141	49	2.9	2.79	-.56	.25	.81	-.9	.81	-.9	1.16	.74	.75	2	O2
144	49	2.9	2.85	-.74	.25	.65	-1.9	.68	-1.7	1.30	.83	.75	11	D8
145	49	3.0	2.86	-.80	.25	.88	-.5	.87	-.6	1.13	.78	.74	5	D2
146	49	3.0	2.88	-.87	.25	1.10	.5	1.09	.4	.83	.67	.74	9	D6
148	49	3.0	2.92	-.99	.25	1.04	.2	1.07	.3	.93	.70	.74	4	D1
151	49	3.1	2.97	-1.18	.25	1.26	1.2	1.22	1.0	.68	.67	.74	10	D7 least difficult
132.0	49.0	2.7	2.55	.00	.25	.98	-.2	.95	-.3		.75			Mean (Count: 14)
20.1	.0	.4	.50	1.25	.00	.29	1.4	.25	1.2		.06			S.D. (Population)
20.9	.0	.4	.52	1.30	.01	.30	1.5	.26	1.3		.07			S.D. (Sample)

Model, Populn: RMSE .25 Adj (True) S.D. 1.22 Separation 4.87 Strata 6.82 Reliability .96
Model, Sample: RMSE .25 Adj (True) S.D. 1.27 Separation 5.06 Strata 7.08 Reliability .96
Model, Fixed (all same) chi-square: 328.5 d.f.: 13 significance (probability): .00
Model, Random (normal) chi-square: 12.5 d.f.: 12 significance (probability): .40

Figure 2. Criteria Measurement Report

The difficulty level ranges from the maximum logit at 2.43 to the minimum logit (least difficult) at -1.18logit with a standard deviation of (S.D.) 1.22. The item separation of 4.87, which can be round of to 5 is different levels of item difficulty. The most difficult criteria is, M=Model, all the M items are located way above the rest of the items, Highest at 2.43 logit and lowest at 2.16. Mean for M=7.02/3=2.34 Where else, the oral and graphic presentation is mediocre, Mean for O= -1.37/3= -0.46 Design development is the easiest with the lowest mean value, Mean for D=-5.64/8=-0.71

Cat	Score	Exp.	Resd	StRes	
2.69	2.69	2.70	.00	.00	Mean (Count: 686)
.95	.95	.76	.56	.98	S.D. (Population)
.95	.95	.76	.56	.98	S.D. (Sample)

Raw-score variance of observations = 0.90 100.00%

Variance explained by Rasch measures = 0.59 **65.26%**

Figure 3. Measurable Data Summary

Raw variance explained by measures is at 65.26% higher than 40% gives an indication that the instrument is measuring one dimension (uni-dimensionality) or in other words, measuring what is supposed to measure.

6. Conclusions

After all, whole recommendations were implemented to studio in three phases which were working period, interim and final submission. After each session students were asked fill up a questionnaire which were included smiley icons in five levels. This format has chosen to be more easy and fast for students to fill. Finally these five turned to liker scale and analyzed. In summary, sessions proved that recommendations increased efficiency of crit sessions. Also to monitor and evaluate the assessment and grading model that were proposed Rasch measurement and mini facet software used. The results from Rasch shown that criteria based assessment model and defined

criteria for course and assignment objectives worked well and was designed properly. About tutor performance Rasch shown that experience has an important role that can affect the judgment. In summary Rasch also confirmed the necessity of implementing some recommendations like holding workshops for tutors, clarification in objectives and criteria for teachers and students.

Acknowledgements

The author gratefully acknowledge to the Centre for Engineering Education Research, Universiti Kebangsaan Malaysia in the effort of improving the quality of teaching and learning in engineering education and appreciate the cooperative among the project partners.

References

Astin. A.W., Banta, T.W., Cross, K.P., El_Kawas, E. Ewell, P.T., Hutchings, P., & Wright, B.D. (1991). 9 Principles of Good Practice for Assessing Student Learning, The American Association for Higher Education. Stylus Publishing, LLC.

Azrilah A. A., Saidfudin, M, NoorHabibah Arshad, Zaharim, A, Azlinah M , Ghulman, H.A. & Sohaimi Z. (2008). Application of Rasch Model in validating the construct of measurement instrument. International Journal of Education and Information Technologies, Vol.2, Issue 2, pp. 105-112. UK: NAUN Press.

Black P.J., & Wiliam D. (1998). Assessment and Classroom Learning, Assessment in Education, (pp.7-74).

Birenbaum & Dochy, F. (2009). Alternatives in Assessment of Achievement, Learning Processes and Prior Knowledge, (Eds). Norwell MA: Kluwer.

Dochy, Filip, J. R. C., & McDowell, L. (1997). Introduction assessment as a tool for learning. *Studies in Educational Evaluation*, 23(4), 279-298, Elsevier Science Ltd.

Gijbels, D., & Dochy, F. (2006). Students' assessment preferences and approaches to learning: can formative assessment make a difference, *Educational Studies*, 32(4).

Hamzah, A. G., & Mohd Saidfudin, M. (2009). Measurement Paradigm in engineering education:Easier to read and better analysis using Rasch-based Approach. *International Conference on Engineering Education (ICEED 2009)*. Kuala Lumpur.

Hussein Mahmood. (1993). Kepimpinan dan Keberkesanan sekolah. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Inbar-Lourie, O. (2008). Language assessment culture. In Shohamy, E. and Hornberger, N. (Eds.), Encyclopedia of Language and Education (Vol. 7, pp. 285-300), Language Testing and Assessment. New York: Springer.

Nitko, J. (2000). Educational Assessment of Students. Upper Saddle River, New Jersey USA: Prentice Hall.

Prosser, M., Trigwell, K., Hazel, E., & Gallagher, P. (1994). Research and Development in HigherEducation, 16, (pp.305-310).

Rozeha, A, R. Razimah, A., Hamzah. A.G., & Mohd Saidfudin, M. (2008). Application of Rasch-based ESPEGS model in measuring generic skills of engineering students: A New Paradigm. *WSEAS Transactions on Advances in Engineering Education, Vol.5, Issue 8*, 591-602. ISSN: 1790-1979.