International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 13 (2017) pp. 3887-3893

Evaluating the Reliability and Quality of Examination Paper for Multi-tier Application Development Course using Rasch Measurement Model

Zuhaira Muhammad Zain

Information Systems Department, College of Computer and Information Sciences,
Princess Nourah Bint Abdulrahman University, Riyadh, KSA.

ORCID: 0000-0002-5973-387X

Abstract

Final exam has been used tremendously as an assessment tool to measure students' academic performance in most of the higher institutions in the Kingdom of Saudi Arabia. A good quality of a set of constructed items/questions on final exam would be able to measure both students' academic performance and their cognitive skills. Rasch Measurement Model has been used to evaluate the reliability and quality of the final exam questions for the Multi-tier Application Development course. The analysis indicated that the reliability and quality of the final exam questions constructed were relatively good and calibrated with students' learned ability.

Keywords: Bloom's Taxanomy; Information systems; item constructions; quality; Rasch Model; reliability; students performance

INTRODUCTION

Nowadays, universities in Saudi Arabia need to comply to the American Accreditation Board of Engineering and Technology, 2000 (ABET) program accreditation requirements. One of the ABET general criteria is student. Student performance must be evaluated to monitor student progress in order to foster success in attaining student outcomes, thereby enabling graduates to attain program educational objectives [1]. Normally, student performance measurement has been essentially dependent on the students' performance in carrying out tasks such as quizzes, assignments, mid examinations, projects and final exams. A quality task should provide the same level of cognitive thinking skills to all students on what they have learned. In order to increase the students' performance quality, well organized and constructed tasks which are based on Bloom's cognitive thinking skills and the level of students' ability should be considered. A reliable and high quality assessment tools in teaching and learning process is required to measure students' understanding and ability.

Multi-tier Application Development (IS333D) is one of new courses introduced in the Information Systems (IS) Department at the College of Computer and Information Sciences (CCIS) at the Princess Nourah Bint Abdulrahman University (PNU). It is one of the core courses that must be completed by the IS students before they can graduate. The main objective of this course is to introduce the concept of multi-tier architecture to the students and they need to apply it in the Web application development.

In this paper, the final examination questions for IS333D for Semester 1 Session 2015/2016 are taken into account as the assessment tool. Furthermore, in the process of constructing these examination questions, it is vital to have fairly distributed examination questions based on Bloom's cognitive thinking skills, the level of students' ability and level of questions/items difficulty. According to Morales, a discussion of reliability is essential in evaluating the quality of the questions [2]. The reliability is the degree to which an instrument consistently measures the ability of an individual or group. Generally, to the best of the author's knowledge, in CCIS, there is no statistical measurement on reliability of any examination questions. The questions were only checked for their format, spelling, and the relevance of questions by the course specialist. Consequently, there is no statistical evidence to verify that a set of examination questions is reliable.

The Rasch Measurement Model has been used to assess the reliability and quality of examination paper of some Engineering courses in Malaysia [3, 4, 5, 6, 7], nevertheless, to the best of the researcher's knowledge, it has not been applied for Information Systems courses especially in Saudi Arabia. The model fulfill the guidelines that has been emphasized by Wright and Mok [8] that a measurement model must produce linear measures, overcome missing data, provide estimates of precision, detect misfits, and distinguish the parameters of the object being measured from those of the measuring instrument. Thus, it can generate meaningful inferences by transforming an ordinal score into a linear, interval-level variable, through estimating the fit of data to the Rasch model's expectations. The basic principle underlying the Rasch Model is that the probability of a respondent/student successfully verifying a particular item/question is governed by the difference between the item/question's difficulty and respondent/student's ability [9, 10, 11]. The logic underlying this principle is that all respondents/students have a higher probability of answering easier items/questions and a lower probability of answering more difficult items/questions accurately [9]. Moreover, Rasch Model is one of the reliable and appropriate method in

assessing students' ability [4, 6, 7]. Aziz et al. specified that the model's Wright Map can give a precise overview of the student's achievement on a linear scale of measurement [12]. Another study by Rashid et al. disclosed that Rasch Model Wright Map could provide meaningful information on the students' learning effectiveness [13].

The purpose of this study is to evaluate the reliability and the quality of final examination questions for IS333D course by using Rasch Measurement Model. The evaluation is to check whether the constructed questions calibrate with students' learning abilities and the course contents or not. It is part of the study to enrich and improve students' cognitive thinking skills and ability in developing multi-tier application.

METHODOLOGY

The data was obtained from the final examination questions of IS333D course, which was taken by the third year Information Systems students of CCIS, PNU. Data from 77 students were collected and studied. The final examination consists of 36 questions which was divided into five parts, Part A, Part B, Part C, Part D and Part E. Students are required to answers all questions. The questions covering four learning topics in IS333D, Principles of Object Oriented Web Programming Language (PHP), HTTP, Multi-tier application development and deployment and Secure the web application. The course learning outcomes for the four learning topics for IS333D expected for the students to achieve is shown in Table I.

Table I: Course learning outcomes for four learning topics for IS333D

No	Course Learning Outcomes
1	Able to identify the basic concept of PHP and use it in the development of multi-tier application.
2	Able to describe the basic of HTTP and use it to pass a user key-in data.
3	Able to implement and deploy a multi-tier software application.
4	Able to identify how to interact with a database management system.
5	Able to identify the basic mechanisms for accessing relational databases from various types of application development environments.
6	Able to design the basics of linking data/ information modeling and business process modeling.
7	Able to identify the basics of securing web applications.

The questions are entered as entry number as shown in Table II. The item is labelled as Question No., and Taxonomy Bloom Level of Learning, which the students expected to develop three Level of Bloom's Taxonomy, namely Remembering/ Understanding (Level 1), Applying/ Analyzing (Level 2), Evaluating/ Creating (Level 3). Thus for entry item number 1, the item is coded as A01_1 (refer to Table II) where A01 is equal to question number and _1 is equal to the Taxonomy Bloom Level of Learning.

Table II: Entry number coded for each exam question

Part	Qs.	Entry No.	Learning Topics
A	1	A01_1	Multi-tier application development and deployment
	2	A02_1	Principles of Object Oriented Web Programming Language (PHP)
	3	A03_1	Secure the web application
	4	A04_1	HTTP
	5	A05_1	HTTP
	6	A06_1	Principles of Object Oriented Web Programming Language (PHP)
	7	A07_1	HTTP
	8	A08_1	Multi-tier application development and deployment
	9	A09_1	Principles of Object Oriented Web Programming Language (PHP)
	10	A10_1	Secure the web application
B	11	B11_1	Principles of Object Oriented Web Programming Language (PHP)
	12	B12_1	Secure the web application
	13	B13_1	Principles of Object Oriented Web Programming Language (PHP)
	14	B14_1	HTTP
	15	B15_1	Principles of Object Oriented Web Programming Language (PHP)
	16	B16_1	Principles of Object Oriented Web Programming Language (PHP)
	17	B17_1	Multi-tier application development and deployment
	18	B18_1	Multi-tier application development and deployment
	19	B19_1	HTTP
	20	B20_2	Principles of Object Oriented Web Programming Language (PHP)
C	i	C21_2	Principles of Object Oriented Web Programming Language (PHP)
	ii	C22_2	
	iii	C23_2	
	iv	C24_2	
	v	C25_2	
D	i	D26_2	Principles of Object Oriented Web

Part	Qs.	Entry No.	Learning Topics
	ii	D27_2	Programming Language (PHP)
	iii	D28_2	
	iv	D29_2	
	v	D30_2	
	vi	D31_2	
	vii	D32_2	
	viii	D33_2	
	ix	D34_2	
	x	D35_2	
E	36	E36_3	Multi-tier application development and deployment

Score from final examination results were gathered and compiled. As these raw score have different total marks for each question, a standardization method is used. The formula for the standardization is given below [14]:

$$z_{ij} = \frac{x_{ij} - min\ x_j}{\max x_j} \quad (1)$$

where i = the ith students (i = 1, 2, ... , 77), j = the jth questions (j = 1, 2, ..., 36), z_{ij} = standardized marks for ith student and jth question, x_{ij} = marks for ith student and jth question, $min\ x_j$ = minimum marks for jth question, and $max\ x_j$ = maximum marks for jth question.

Responses from the students' exam results were analysed using rating scale in which the students were rated according to their achievement. From (1),

$$z_{ij} * 10 = A \quad (2)$$

Then, A is classified correspond to the rating scale in Table III:

Table III: Marks (A) and corresponding rating scales

Marks (*A*)	0-1.49	1.50-3.49	3.50-6.49	6.50-8.49	8.50-10
Rating scale	1	2	3	4	5

This grade rating is tabulated in Excel*prn format. Using Rasch software, Bond & Fox Steps, this numerical coding is necessary for further evaluation of the students' achievement and also the reliability and the quality of items. The analysis outputs obtained from the Bond & Fox Steps were analysed and studied.

RESULTS AND DISCUSSIONS

Fig. 1 illustrates the Summary Statistics for 36 IS333D final exam questions answered by 77 students. The person's mean of +.68 (SE .17) is the first indicator that the students find this set of final exam questions comparatively easy. This means that they tend to answer all the questions correctly. The mean square fit (IMNSQ and OMNSQ) and the z statistic (INFIT ZSTD and OUTFIT ZSTD) are closer to their expected values, +1 and 0 respectively for items and persons. This shows satisfactory fit to the model. Moreover, the item reliability (Rasch equivalence to Cronbach's alpha) is .85 while person reliability is much lower at .73. The values of item and person reliability (> 0.6) do confirm that the instrument used for measuring the student learning ability for IS333D is reliable, reproducible, and valid for measurement.

Fig. 2 presents the variable-map representation for the analysis. It shows that the distribution of the students is on the left and the distribution of the questions/items is on the right according to person and item label respectively. The result from the summary statistics (see Fig. 1) was supported by the person and item distributions.

Fig. 3 demonstrated a segment of the output of the Rasch item estimates for the IS333D final exam questions, so the detail of the variable-map locations can be verified more easily. Items are represented by the Entry Number (see Table 2). The easiest question/item is D32_2 located at the bottom of the item distribution at –2.37 logits (SE 1.46), while the most difficult question/item is A10_1 located at +1.01 logits (SE .08). The analysis shows that the easiest question (D32_2) has minimum estimated measure. This means that the question can be answered correctly by all students. The fit statistics of the questions/items output look good, although we need to reconsider two under fit questions/items, D34_2 and A08_1. Counter check against the Guttman scalogram (see Fig. 4) indicates that the two questions/items, D34_2 (item 34) and A08_1 (item 8) have been under rated by two of the top 11 students, student 8 (R08) and student 31 (R31) respectively. One possible reason is that they could have been careless in attempting their answers which led to such a grossly under rated work. After verifying that the Point Measure Correlation (see PTMEA CORR column in Fig. 3) for both questions/items are positive value, the two misfits are acceptable.

The principal contrast analysis of the Rasch residual variance is shown in Fig. 5. The variance explained by measures is good (61.2%). The uni-dimensionality of the final examination instrument is strongly confirmed by having a good unexplained variance in the first contrast (3.8%). Hence, it proved that the final examination questions are only related with the content of the IS333D course.

Persons	77 INPUT		77 MEASURED			INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	ERROR		IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	148.7	35.0	.68	.17		1.02	.1	.99	.1
S.D.	17.1	.0	.36	.08		.25	.8	.51	.8
REAL RMSE	.19	ADJ.SD	.31	SEPARATION	1.63	Person RELIABILITY		.73	
Items	36 INPUT		36 MEASURED			INFIT		OUTFIT	
MEAN	318.6	75.0	.00	.13		1.09	-.1	.99	-.1
S.D.	41.7	.0	.41	.09		.35	2.5	.51	1.5
REAL RMSE	.16	ADJ.SD	.38	SEPARATION	2.42	Item RELIABILITY		.85	

Figure 1: Summary statistics

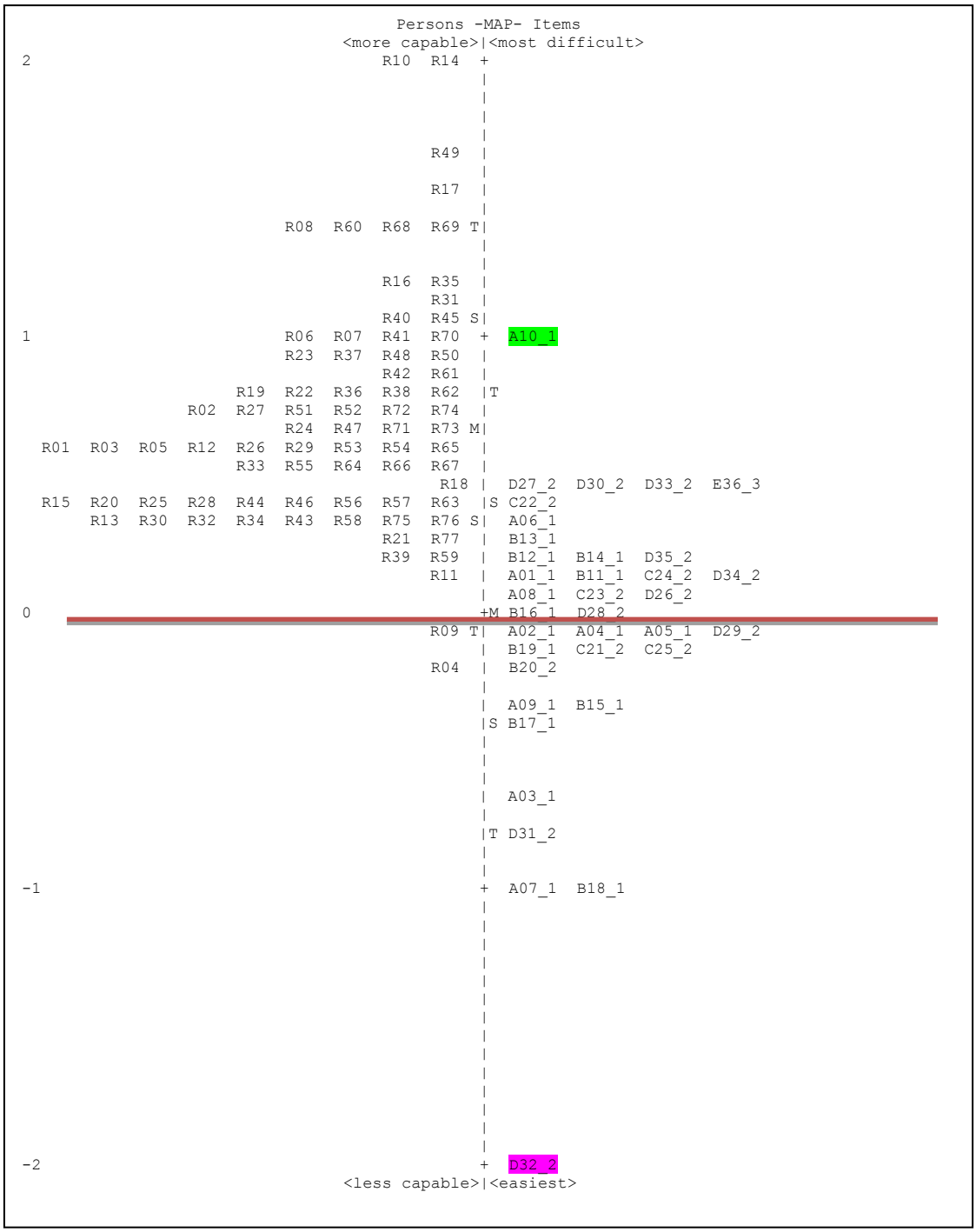

Figure 2: Variable map

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
10	163	75	1.01	.08	1.17	1.1	1.03	.2	.54	30.7	30.6	A10_1
30	257	75	.50	.07	1.11	1.0	.97	-.1	.47	13.3	18.5	D30_2
36	257	75	.50	.07	.28	-9.2	.36	-4.3	.61	46.7	18.5	E36_3
33	263	75	.46	.07	.38	-7.1	.41	-3.7	.54	41.3	18.4	D33_2
27	267	75	.44	.07	1.27	2.1	1.24	1.1	.38	14.7	20.5	D27_2
22	280	75	.37	.08	.56	-4.1	.53	-2.4	.53	29.3	22.8	C22_2
6	283	75	.35	.08	1.19	1.4	1.29	1.2	.38	20.0	25.3	A06_1
13	295	75	.28	.08	1.24	1.6	1.55	1.8	.34	25.3	28.6	B13_1
12	303	75	.23	.08	1.05	.4	.84	-.5	.44	29.3	30.0	B12_1
35	303	75	.23	.08	.72	-2.0	.67	-1.2	.42	34.7	30.0	D35_2
14	311	75	.17	.09	1.03	.3	1.00	.1	.42	32.0	33.3	B14_1
24	312	75	.16	.09	.67	-2.1	1.05	.3	.26	29.3	33.0	C24_2
1	315	75	.14	.09	1.18	1.0	.99	.1	.35	38.7	36.2	A01_1
34	315	75	.14	.09	1.56	2.8	2.53	3.4	.12	34.7	36.2	D34_2
11	319	75	.11	.09	1.39	1.9	1.12	.4	.27	38.7	37.6	B11_1
23	321	75	.09	.09	.63	-2.1	.61	-1.2	.35	41.3	39.9	C23_2
8	323	75	.08	.09	1.64	2.8	2.42	3.0	.08	38.7	42.2	A08_1
26	327	75	.04	.10	1.38	1.7	1.36	1.0	.22	44.0	44.3	D26_2
16	331	75	.01	.10	1.08	.4	.75	-.5	.38	52.0	47.9	B16_1
28	331	75	.01	.10	.94	-.2	.98	.1	.28	45.3	47.9	D28_2
4	335	75	-.04	.10	1.19	.8	1.01	.2	.31	56.0	54.7	A04_1
29	337	75	-.06	.10	.70	-1.2	.69	-.7	.31	54.7	57.1	D29_2
2	339	75	-.08	.11	1.19	.8	.92	.0	.31	64.0	59.7	A02_1
5	339	75	-.08	.11	1.21	.8	1.00	.1	.30	61.3	59.7	A05_1
19	343	75	-.13	.11	1.10	.4	.71	-.5	.35	66.7	62.5	B19_1
21	344	75	-.14	.11	.67	-1.2	.78	-.3	.25	56.0	62.7	C21_2
25	345	75	-.15	.12	.56	-1.7	.85	-.1	.25	57.3	62.9	C25_2
20	347	75	-.18	.12	1.27	.9	1.00	.2	.27	68.0	64.3	B20_2
9	355	75	-.32	.14	1.23	.7	1.48	.9	.22	88.0	84.4	A09_1
15	355	75	-.32	.14	1.36	1.0	1.80	1.3	.19	88.0	84.4	B15_1
17	359	75	-.40	.16	1.53	1.2	1.14	.4	.15	90.7	88.8	B17_1
3	367	75	-.68	.23	1.24	.6	.25	-1.0	.30	97.3	94.2	A03_1
31	369	75	-.80	.27	1.15	.4	.33	-.6	.24	97.3	95.2	D31_2
7	371	75	-.98	.33	1.53	.8	.23	-.7	.23	98.7	97.3	A07_1
18	371	75	-.98	.33	1.70	.9	.61	.0	.15	98.7	97.3	B18_1
32	375	75	-2.37	1.46	MINIMUM ESTIMATED MEASURE							D32_2
MEAN	320.2	75.0	-.07	.16	1.09	-.1	.99	-.1		52.1	50.5	
S.D.	42.1	.0	.56	.23	.35	2.5	.51	1.5		25.0	24.7	

Figure 3. Item measure
Note: Acceptable range for Infit and Outfit Mean-square is between 0.6 to 1.4 [15] and acceptable range for Infit and Outfit Z-std is between -2 to +2 [9].

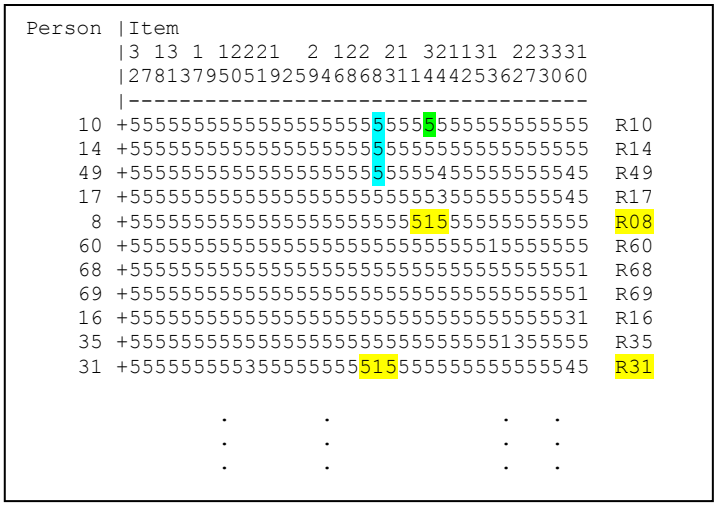

Figure 4. Guttman scalogram

STANDARDIZED RESIDUAL VARIANCE SCREE PLOT
Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

		Empirical			Modeled
Total variance in observations	=	90.3	100.0%		100.0%
Variance explained by measures	=	55.3	61.2%		60.0%
Unexplained variance (total)	=	35.0	38.8%	100.0%	40.0%
Unexplned variance in 1st contrast	=	3.4	3.8%	9.7%	
Unexplned variance in 2nd contrast	=	2.6	2.8%	7.3%	
Unexplned variance in 3rd contrast	=	2.2	2.4%	6.3%	
Unexplned variance in 4th contrast	=	1.8	2.0%	5.2%	

Figure 5. Principal contrast analysis
Note: Variance explained by measures should be ≥ 50% and unexplained variance in the first contrast should be ≤ 15% [16].

CONCLUSIONS AND FUTURE WORKS

This paper explained the evaluation of reliability and quality of final examination paper for Multi-tier Application Development (IS333D) course for IS students at Princess Nourah Bint Abdulrahman University by using Rasch Measurement Model.

In conclusions, this study confirms the reliability and quality of the 36 items/questions of the final examination paper for IS333D via a reliable and valid analysis. The constructed items/questions are reliable and in high quality to measure the students' academic performance. This findings can be the future references for items/questions development of other Information Systems courses.

It is also found that, although the small sample size is used, Rasch Measurement Model is an effective tool in assessing the reliability and quality of final examination paper accurately and fast by classifying the questions according to students' learning ability and their cognitive thinking skills. Hence, the model allows each question (item) to be evaluated discretely and calibrated with students' learning abilities and the course contents. It also accurately categorized the students according to their capability to answer the questions.

In the future, we will continue our efforts to evaluate the reliability and quality for other assessment tools for other IS courses in order to prepare the evidences for the quality of student performance measurement. This will be a great asset for the IS department in complying with the ABET accreditation requirement.

REFERENCES

[1] A. A. Aziz, N. Khatimin, A. Zaharim, and T. S. A. Salleh, "Evaluating multiple choice items in determining quality of test", Teaching Assessment and Learning for Engineering (TALE), pp. 565–569, 2013.

[2] R.A. Morales, "Evaluation of Mathematics achievement test: A comparison between CTT and IRT," The International Journal Educational and Psychological Assessment, vol.1, pp.19-26, 2009.

[3] H. Othman, I. Asshaari, H. Bahaludin, Z. M. Nopiah, and N. A. Ismail, "Application of Rasch measurement model in reliability and quality evaluation of examination paper for Engineering Mathematics courses", Procedia Social and Behavioral Sciences, vol. 60, pp. 163-171, 2012.

[4] R. F. M. Said, "Application of Rasch measurement model in evaluating student performance for Foundation of Computing II", in 7th International Conference on University Learning and Teaching (InCULT 2014) Proceedings, Springer Singapore, pp. 251-259, 2016.

[5] H. Othman, N. A. Ismail, I. Asshaari, F. M. Hamzah, and Z. M. Nopiah, "Application of Rasch measurement model for reliability measurement instrument in Vector Calculus course", Journal of Engineering Science and Technology, vol. 10, no. 2, pp. 77-83, 2015.

[6] A. A. Aziz, A. Zaharim, N. F. A. Fuaad, and Z. M. Nopiah, "Students' performance on engineering mathematics: Applying Rasch measurement model", Information Technology Based Higher Education and Training (ITHET), pp. 1–4, 2013.

[7] S. A. Osman, M. A. Khoiry, W. H. W. Badaruzzaman, and A. Mutalib, "Measurement of students' understanding in final examination of statics and dynamics course using Rasch measurement model", Teaching Assessment and Learning for Engineering (TALE), pp. 805–810, 2013.

[8] B. D. Wright and M. M. C. Mok, "An overview of the family of Rasch measurement models", in E. V. Smith & R. M. Smith (Eds.), Introduction to Rasch Measurement, Maple Grove, MN: JAM Press, pp. 1-24, 2004.

[9] T.V. Bond and C.M. Fox, "Applying the Rasch model: Fundamental measurement in the Human Sciences", 2nd ed., New Jersey: Lawrence Erlbaum Associates, 2008.

[10] G. Rasch, "Probabilistic models for some intelligence and attainment test", University of Chicago Press, Chicago, 1960.

[11] B. D. Wright and M. H. Stone, "Best test design". Chicago, IL: MESA Press, 1979.

[12] A.A. Aziz, A. Mohamed, N.H. Arshad, S. Zakaria, and S. Masodi, "Appraisal on course learning outcomes using Rasch measurement: A case study in Information Technology education," International Journal of Systems Application, Engineering and Development, pp.164–171, 2007.

[13] A.R. Rashid, A. Zaharim, and S. Masodi, "Application of Rasch measurement in evaluation of learning outcome: A case study in Electrical Engineering", Regional Conference on Engineering Mathematics, Mechanics, Manufacturing & Architecture (EMARC), pp.151-165, 2007.

[14] H. Othman, I. Asshaari, H. Bahaludin, Z.M. Nopiah, and N.A. Ismail, "Application of Rasch measurement model in reliability and quality evaluations of examination paper for Engineering Mathematics courses," Procedia Social and Behavioral Sciences, vol.60, pp. 163-171, 2012.

[15] B.D. Wright, M. Linacre, J.-E. Gustafsson, and P. Martin-Loff, "Reasonable mean-square fit values", Rasch Measurement Transactions, vol.3, pp.370, 1994.

[16] W.P.Jr. Fisher, "Rating scale instrument quality criteria," Rasch Measurement Transactions, vol.21, pp.1095, 2007.