International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:8, No:4, 2014

Personalized Learning: An Analysis Using Item Response Theory

A. Yacob, N. Hj. Ali, M. H. Yusoff, M. Y. Mohd Saman, W. M. A. F. W. Hamzah

***Abstract*—**Personalized learning becomes increasingly popular which not be restricted by time, place or any other barriers. This study proposes an analysis of Personalized Learning using Item Response Theory which considers course material difficulty and learner ability. The study investigates twenty undergraduate students at TATI University College, who are taking programming subject. By using the IRT, it was found that, finding the most appropriate problem levels to each student include high and low level test items together is not a problem. Thus, the student abilities can be asses more accurately and fairly. Learners who experience more anxiety will affect a heavier cognitive load and receive lower test scores. Instructors are encouraged to provide a supportive learning environment to enhance learning effectiveness because Cognitive Load Theory concerns the limited capacity of the brain to absorb new information.

***Keywords*—**Analysis, Cognitive Load Theory, Item Response Theory, Learning, Motivation, Performance.

I. Introduction

THE current trend in education emphasizes on identifying just-in-time delivering method, tailored differences in learners' skills level, perspectives, culture and other educational contexts. Besides, rapid evolution of ICT enabled technological tools for facilitating the implementation of the new paradigm in education. Personalized learning should be tailored to the continuously modified individual learner's requirements, abilities, preferences, background knowledge, interests and skills.

Powerful methods for data collection like tests, surveys and questionnaires much used for research purposes. However, the mentioned methods are not an easy undertaking. Item Response Theory (IRT) is a paradigm for the design, analysis, and scoring of tests, questionnaires, and similar instruments measuring abilities, attitudes, or other variables. IRT also known as latent trait theory, strong true score theory, or modern mental test theory. The term item covers all kinds of informative item including multiple-choice questions or likert scale question. IRT is used to evaluate the performance of items and sets of items. This feature of IRT is very useful in constructing short forms used to ensure that the items provide adequate precision across the entire range of interest.

A. Yacob is with the Faculty Computer, Media & Technology Management, TATI University College, 24000 Kemaman, Terengganu (e-mail: azliza@tatiuc.edu.my).

N. H. Ali, M. H. Yusoff, M. Y. Mohd Saman, and W. M. Amir Fazamin W. Hamzah are with the Department of Computer Science, UMT, 21030, Kuala Terengganu, Terengganu (e-mail: aida@umt.edu.my, hafiz.yusoff@umt.edu.my, yazid@umt.edu.my, amirfazamin@gmail.com).

In this research, IRT is applied into the programming learning system. By simply dragging 0 or 1 scores from the test result saved in Excel file to IRTs' software, teachers can obtain students' abilities and parameters for each problem attached in the learning system. The proposed learning system based on IRT provides benefits in providing learning paths that can be adapted to various levels of item difficulty and various learners' abilities. Learning guidelines that provided in the system can prevent learners becoming lost in the course materials. Thus, cognitive loading also can be reduced by filtering unsuitable course materials.

This work is organized as follow: Introduction in Section I. Section II presents the Literature Review, including Item Response Theory Model, Fit Statistics, Separation Value, IRT software and Cognitive Load Theory. Section III will discuss about Methodology, and Implementation of IRTin Section IV. Meanwhile Results and Discussions will discuss in Section V. Finally Conclusion and Future Works will be presented in Section VI.

II. Literature Review

Item Response Theory usually is applied in the Computerized Adaptive Test (CAT) domain, to select the most appropriate items for examinees based on individual ability [1], [2]. Traditional measurement instruments such as paper–pencil tests and fixed-content is replaced by the CAT for the sake of efficiency and effectiveness [3], [4]. Rasch model referred to as a prescriptive model which prescribes specific conditions for the data to meet. Rasch Analysis is commonly used to analyze test data and Likert survey data, to construct and evaluate question item banks, and to evaluate development [5]. The whole research process must be in line with the model's specifications starting from the beginning.

Because of their general applicability, they are increasingly being used not just in psychometrics, but in health profession, marketing and education fields.

A. Item Response Theory Model

IRT describe the relationship between a respondent's answer to a survey question and respondent's level of the 'latent variable' (θ), being measured by the scale using a set of mathematical models. IRT model purposely used to estimate an examinee's ability (θ) or proficiency according to their dichotomous responses (true/false) to test items [6]. Item difficulty is estimated separately for each item [7]. According to the IRT model, item characteristic curve (ICC) were results from the relationship between examinee's responses and test items [8]. ICC is the basic of IRT which all the other

constructs of the theory depend upon this curve. According to [9], major benefits of IRT are its comprehensive representation of content, ant its ability to determine the optimal number of response number of response categories for individual items. Fig. 1 represents the sample item characteristic curve (ICC). The horizontal axis is used for ability scale; meanwhile the vertical axis is for the probability that an examinee with certain ability will give a correct answer to the item. If the slope is large, thus the curve is steeper, it indicate that students of high abilities have a greater probability of correct response [10]. As mentioned by [6], this probability will be smaller for examinees of low ability and larger for examinees of high ability.

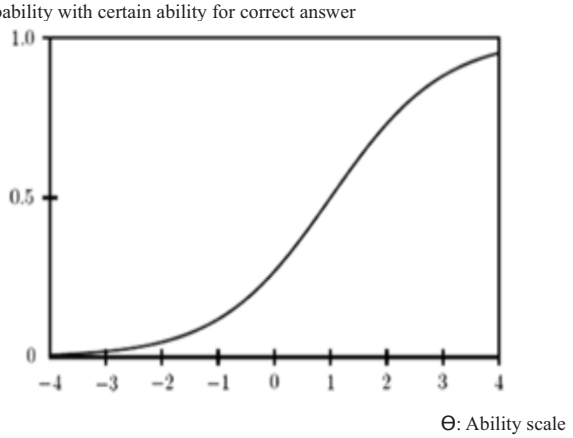


Fig. 1 Sample item characteristic curve (ICC)

Before reasonable and precise interpretations based on IRT can be made, several assumptions must be met: Unidimensionality assumption and Local independence assumption. Rasch analysis is assumed as independent of items, that is the probability of answer an item correctly is not depending on the answer to other questions [11]. It is not suitable for tests where the questions are structured into parts, (the latter parts relying on the earlier ones). According to [5], the dichotomous Rasch Model, makes the further assumption that the probability of such a person answering a particular item correctly is a function of the person's ability and the item's difficulty alone.

Although evaluation for each assumptions is important, IRT models are robust to minor violations and no real data ever meet the assumptions perfectly [9]. Although IRT requires extremely large sample sizes, these theory offers a powerful tool to evaluate the precision of questionnaire measurement [12], [13]. Sample sizes as small as 100 are often adequate for estimating stable Rasch-model parameters [13].

The Rasch Model is a psychometric model, used for analyzing categorical data. For the purpose of assessments evaluation, Rasch model has been used to estimates item-difficulty and person-ability that are not dependent on raw scores. Rasch analysis is based on a prescriptive model rather than being pre-determined, which has replaced classical test theory (CTT) [14].Results from the analysis can be used to identify items which are causing unexpected response and require for some modification[14], [5]. Rasch analysis also frequently used toprovide validity evidence for pedagogical assessments [15], [16]. Table I contains the terminology used in IRT model.

TABLE I
TERMS USED IN IRT MODEL [9]

Terminology	Description
Scale	Multiple items that measure a single domain such as fatigue.
Item	A question in a scale.
Theta (θ)	Unobservable construct (or latent variable) being measured by a scale.
Classical test theory (CTT)	Traditional psychometric methods such as factor analysis and Cronbach's α, in contrast to IRT
Item characteristic curve (ICC) or Category Response Curve (CRC)	Models the probabilistic relationship between a person's response to each category for an item and their level on the underlying construct (θ).
Discrimination parameter (a, α) – slope	IRT model item parameter that indicates the strength of the relationship between an item and the measured construct. Also, the parameter indicates how well an item discriminates between respondents below and above the item threshold parameter, as indicated by the slope of the ICCs.
Dichotomous response categories	Contain two response categories. Eg: yes/no, true/false, or agree/disagree
Polytomous response categories	An item having more than two response categories. For example, a 5-point Likert type scale.
Threshold parameter (b, β) – difficulty, location	IRT model item parameter that indicates the severity or difficulty of an item response. The location along the θ-continuum of the item response categories.
Unidimensionality assumption	Assumes that only one factor affecting the person's test performance.

According to [17], CTT often referred as weak model because of assumptions that these model are fairly easy met by test data. Meanwhile IRT referred as strong model and assumed that less likely to be met with test data. An important distinction between IRT and CTT is that IRT defines a scale for the underlying latent variable that is being measured by a set of items, and items are calibrated with respect to this same scale [13]. Another advantage of IRT over CTT is that the more sophisticated information IRT provides allows a researcher to improve the reliability of an assessment. A major limitation of CTT is that person ability and item difficulty cannot be estimated separately [18] and the use of Likert scoring with the erroneous allocation of equal weight to all the item in the questionnaire, treating the whole questionnaire as interval scale based on ordinal level scoring [19]. All the mentioned limitations can be overcome by the use of IRT.

Dichomous items are including 1PL, 2PL and 3PL models which contain two response categories such as true/false, yes/no or agree/disagree. From the other side, polytomous item consists of Partial Credit Model (PCM), Rating Scale Model (RSM), Graded Response Model (GRM) and Generalized Partial Credit Model (GPCM) for ordered responses, meanwhile Nominal Model used for items with non-specified response order [13]. This type of item has more than two responses categories, for example, a 5-point Likert type scale.

1 Parameter Logistic (1PL) model, also known as the dichotomous Rasch model [20], [21]. Each item i is characterized by only one parameter, the item difficulty (b_i), in a logistic formation as shown, where D is a scaling factor, whose value 1.7. D is used to change this logistic model equivalent into the normal ogive model, from which the IRT was originally developed from. θ is ability scale. $P_i(\theta)$ indicates the probability of a correct response on item I by a test taker whose ability isθ.

$$P_i(\theta) = \frac{1}{1+\exp(-D(\theta-b_i))} \quad (1)$$

As stated by [13], the two parameter logistic (2PL) model is often applied for the items with dichotomous response options. In the 2PL model, two parameters involved: difficulty (b_i) and discrimination ((a_i), as shown below:

$$P_i(\theta) = \frac{1}{1+\exp(-a_iD(\theta-b_i))} \quad (2)$$

The last 3PL model which includes three item parameters, estimates difficulty (b_i), discrimination (a_i)), and includes the potential guess degree (c_i).

$$P_i(\theta) = c_i + frac1 + c_i1 + \exp(-a_iD(\theta - b_i))$$

or

$$P_i(\theta) = c_i + \frac{(1-c_i)}{1+\exp(-a_iD(\theta-b_i))} \quad (3)$$

The discrimination (a_i) power is graphically represented as the steepness of the curve of ICC. The curve becomes steeper as the discrimination (a_i) parameter increases. The curve of an item with lower value of the discrimination parameter becomes flat. Such items have less power of discriminating the ability of test takers and are useless in the test. The curve of ICC shifts from left to right as the b-parameter gets higher, or the item becomes more difficult. If the c-parameter is set to 0 as in a Two Parameter Model, the central point, or inflection point of the curve is at p=0.5. It means that a person whose ability is = 0 has 0.5 chances of answering the item correctly. Refer to Fig. 2.

1. Fit Statistics

There are two fit statistics that were used, including Infit and Outfit. Fit statistics are calculated, to check for any items that cause unexpected response patterns. References [14], [19] stated that infit statistics reflect to the response patterns where the test is targeting ability while outfit statistics highlight unexpected responses which is more sensitive to outliers. Both of infit and outfit statistics are manifested in mean-square values' (MNSQ) size and z-standardized scores (ZSTD) which indicate the significance of the misfit. Acceptable values for low-stakes multiple choice tests for MNSQs range from 0.7 to 1.3 and -2.0-2.0 for ZSTDs [16], [22]. Misfitting item indicates that the item is either poorly defined or is measuring something different. The item is suggested to be improved or deleted if the item is still misfit. While deleting, it is more important to delete the underfitting ones (>1.3) [19].

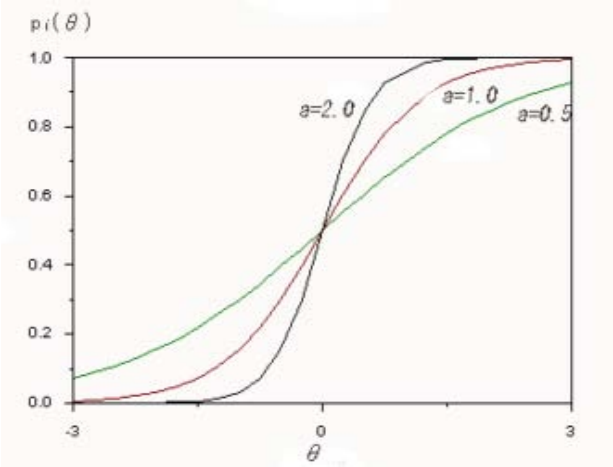


Fig. 2 Shift of ICC according to discrimination parameters

2. Separation Value

Person separation statistics are the key indicators of questionnaire functioning within the Rasch model. Separation indices include the person separation index and reliability. Item strata that labels as Separation from the Summary Statistics also important to investigate the representativeness of the items. According to [16], [19], [23], the minimum value of item strata is 2.0, and it results in a reliability of >0.8 (the minimum recommended acceptance level). It means that if the separation value is higher than 2, it can be said that one can rely on the representativeness of the test items which is acceptable index. The higher the reliability, the better the questionnaire is in terms of its ability to discriminate among subjects.

3. IRT Software

Softwares that can be used for performance measure including BILOG-MG, Facets, ICL, jMetrik, PARSCALE, MULTILOG, STUIRT, Winsteps and many more. Every of the mentioned softwares are designed using different specifications but for the same purpose. Software likes ConQuest 3, FACET, Winsteps, WINMIRA and T-Rasch are some of commercial Rasch Software which needed for licence. Meanwhile Bigsteps,Facets-DOS, Ministep, jMetrik are some of freeware Rasch Software. While the others like BILOG-MG and PARSCALE are paid software (IRT programs with Rasch-like capability). Based on previous study, author found that Winsteps software is the most popular software used to evaluate the performance of item and person. Winsteps are a graphical representation designed for analysis with the Rasch model that enables the researcher to decide if usage is satisfactory or whether changes need to be made. These model referred to as a prescriptive model rather than descriptive because it prescribes specific conditions for the data to meet [16]. It is commercially available from Winsteps and BIGSTEPS, a previous DOS-based version. Winsteps are

used for the response categories that enables the researcher to decide if usage is satisfactory or whether changes need to be made [19].

4. Cognitive Load Theory

Although Cognitive Load Theory (CLT) has been extensively discussed in relation to instructional material design, empirical studies on the relationship between cognitive load, level of item difficulty and contents of the learning system are still few in number. Cognitive load refers to the total amount of mental activity performed by working memory at any point in time [24] which consists of intrinsic, extraneous and germane load. In order to avoid burdening the capacity of working memory unnecessarily, reducing the attentional requirements are suggested [25]. CLT concerns the limited capacity of the brain to absorb new information [26]. According to [24], memory overload is one factor that reported as impeding learning performance. This theory mainly concerned with the development of techniques for managing Working Memory load imposed by a learning task. According to [27], intrinsic cognitive load consist with the elements in learning task which contributed in interaction. Meanwhile, extraneous load is imposed by information and activities that do not directly contribute to learning. On the other hand, germane load is caused by information and activities that foster learning processes. All types of mentioned load are important to realize that the total cognitive load regarding to instructional design should not exceed the available Working Memory processing capacity.

In order to apply total cognitive load, the first step to do is eliminate an extraneous load. According to [28], cueing is expected to reduce extraneous cognitive load associated with locating relevant information. It was revealed by[25], which extraneous cognitive load was reduced by cueing after three exposures. It then, follows by managing an intrinsic and germane load. Simplifying the task in an instructional design is a way to manage an intrinsic load. Meanwhile increasing the variability of learning task is an effective way to increase germane load.

III. Methodology

A. Participants

Twenty undergraduate programming students at TATI University College participated in the present study. Their age ranged from 19-25 who has taken programming subject for the first time (novice).

B. Instruments

The Programming Learning system was developed to provide a conducive and centralized learning system to novice learners. The instrument used was 24 items with four multiple choice programming test. Learners were required to choose the best possible answer for each item. Four multiple choices question for programming test is chosen, to be given to the participants, as author knows that they are a type of novice learner. Time allowed for answering all the items was 25 minutes. The data were analyzed using MINISTEP 3.80.1, under WINSTEPS software. The present research employed with four chapters which each of the chapter contain with six questions. The four chapters that involved are: chapter A for Basic of C++, chapter B for Input Output, chapter C for Selection, and chapter D for Repetition. The results from the system then will save to excel and IRTs' software to evaluate the difficulty of items and students' ability.

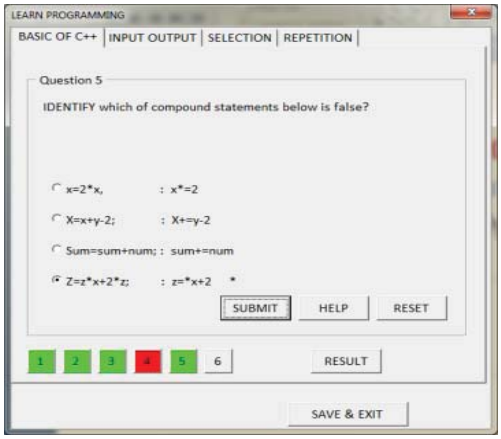
IV. Implementations of IRT

Through readings, it was found that Item Response Theory has been implemented in various fields using suitable software likes Excel, Winstep, Multilog, PARSCALE, Bigsteps and CITAS. The implementation of IRT in education, especially on Mathematic [5], [29]-[30], English [14]-[16], [23], [4] and Chemistry [31] subject using both dichotomous and Polytomous response categories. It also was implemented for the test evaluation using web [32], for questionnaire evaluation and development [9], and also for vocabulary size test validation [15], [16].

TABLE II
Implementations of IRT

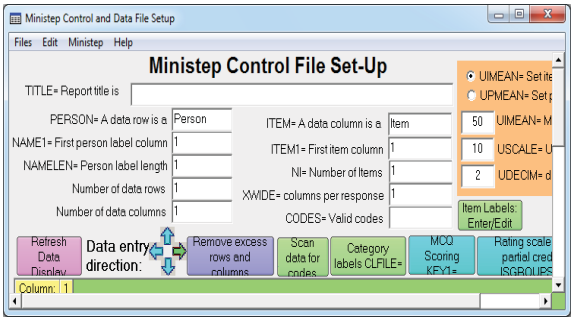
	Fields				Software					
Authors' references	Chemistry	Mathematics	English	Health	Excel	Winstep	Multilog	PARSCALE	CITAS	Bigsteps
[32]		√			√					
[9]				√			√			
[15]			√			√				
[16]			√			√				
[14]			√			√				
[5]		√				√				
[19]				√		√				
[29]		√							√	
[23]			√			√				
[33]		√						√		
[4]			√					√		
[31]	√					√				
[30]		√								√

In learning, [32] found that using IRT, lecturer do not worry about finding the most appropriate problem levels to each student. It means that lecturer can include either high and low level test items together to accurately assess the student abilities. This study also found that most of students are satisfied with this way of testing, which not raises claims about their scores. As stated by [13], benefit of IRT in questionnaire development, evaluation, and refinement, resulting in precise, valid, and relatively brief instruments that minimize response burden. IRT also have been implemented into Web-based learning which achieve personalized learning and help learners to learn more effectively and efficiently[3]. Refer to Table II for the implementations of IRT.



(Learning System)

(Excel File)



(Ministep Software)

Fig. 3 Flow of implementations

In clinical assessment, IRT usage effect to include comprehensive analyses and reduction of measurement error, creation of computer adaptive tests, meaningful scaling of latent variables, objective calibration and equating, evaluation of test and item bias, greater accuracy in the assessment of change due to therapeutic intervention, and evaluation of model and person fit [7]. Fig. 3 is about the flow of implementations.

V. Results and Discussions

A. Variable Maps

A person-item map is a graphical representation that illustrates gaps in person's abilities and item difficulties. Fig. 4 represents the person-item map which can serve to provide evidence for the representativeness of the test items. People called it as, content validity. Items on the right are matched to the persons on the left. The mentioned map able to prove that, the test is appropriately targeted for the group of participants.

According to the Fig. 4, it can be said that item A1 is the easiest item to the participant. Meanwhile item D6 is the most difficult items in this test. The maps also illustrate that there are a gaps between the most difficult items (D6) and the rest items. In this case, the top item may require further investigation either student unfamiliar with the item or confusing happened, or misleading or maybe the question given really the hardest question. Because of the mentioned factors, lecturer should decide the suitability of this item, either to omitted or revise it.

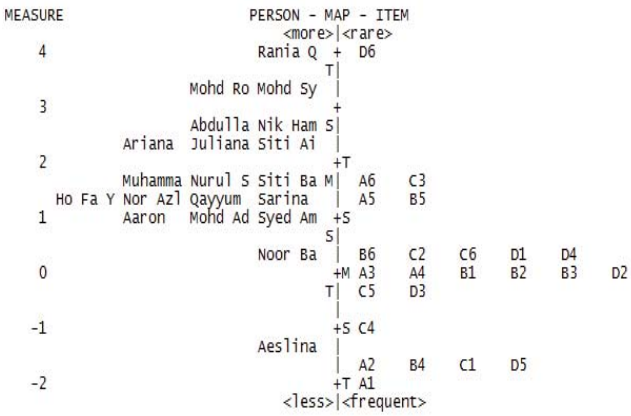


Fig. 4 Variable maps

B. Item Measure

Table III shows the items measure, arranged in ascending order, starting from the most difficult item(D6)to the easiest item (A1). The first column shows the entry number for the system, and the second column for total score. From all twenty participants who have attempted item D6, nobody could get it right. The difficulty of these items is estimated to be 6.26 with the standard error of 1.85; meanwhile the measure value for the easiest item (A1) is -3.08. Item A6 got total score 10 from all 20 participants. The third difficult item in this system is item C3 with the difficulty 1.56. Meanwhile items A5 and B5 got the total score 12 with the difficulty 1.30. 8 easiest items among participants with the difficulty below 1.0 are including items C5, D3, C4, A2, B4, C1, D5, A1.

TABLE III
ITEM MEASURE

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	ITEM
24	0	20	6.26	1.85	D6
6	10	20	1.82	0.51	A6
15	11	20	1.56	0.51	C3
5	12	20	1.30	0.51	A5
11	12	20	1.30	0.51	B5
12	15	20	0.44	0.57	B6
14	15	20	0.44	0.57	C2
18	15	20	0.44	0.57	C6
19	15	20	0.44	0.57	D1
22	15	20	0.44	0.57	D4
3	16	20	0.08	0.62	A3
4	16	20	0.08	0.62	A4
7	16	20	0.08	0.62	B1
8	16	20	0.08	0.62	B2
9	16	20	0.08	0.62	B3
20	16	20	0.08	0.62	D2
17	17	20	-0.34	0.69	C5
21	17	20	-0.34	0.69	D3
16	18	20	-0.90	0.81	C4
2	19	20	-1.77	1.09	A2
10	19	20	-1.77	1.09	B4
13	19	20	-1.77	1.09	C1
23	19	20	-1.77	1.09	D5
1	20	20	-3.08	1.85	A1
MEAN	15.2	20.0	0.13	0.79	
S.D.	4.1	0.0	1.73	0.37	

C. Person Measure

According to Table IV, 8 of the students are located above the 1.81 logit of Person Mean, indicating that they successfully meet the expected performance of the test.

TABLE IV
PERSON MEASURE

ENTRY NUM.	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	PERSON
20	23	24	4.70	1.82	RANIA QISTINA
4	22	24	3.47	1.05	MOHD SYARILL
13	22	24	3.47	1.05	MOHD ROSLAN
12	21	24	2.68	0.77	ABDULLAH
14	21	24	2.68	0.77	NIK HAMDAN
1	20	24	2.18	0.65	ARIANA
11	20	24	2.18	0.65	JULIANA
18	20	24	2.18	0.65	SITI AISYAH
5	19	24	1.80	0.59	NURUL SYUHADA
8	19	24	1.80	0.59	SITI BALQIS
19	19	24	1.80	0.59	MUHAMMAD ALI
9	18	24	1.48	0.55	HO FA YIN
15	18	24	1.48	0.55	SARINA AISYA
3	17	24	1.20	0.52	NOR AZLINDA
10	17	24	1.20	0.52	QAYYUM
2	16	24	0.94	0.50	SYED AMIR
7	16	24	0.94	0.50	AARON
17	16	24	0.94	0.50	MOHD ADLI
16	13	24	0.25	0.47	NOOR BAHIYAH
6	7	24	-1.18	0.53	AESLINA
MEAN	18.2	24.0	1.81	0.69	
S.D	3.5	0.0	1.23	0.31	

The students' names are Rania Qistina, Mohd Syarill, Mohd Roslan, Abdullah, Nik Hamdan, Ariana, Juliana and Siti Aisyah. While the rest participants, Nurul Syuhada, Siti Balqis, Muhammad Ali, Ho Fan Yin, Sarina Aisya, Nor Azlinda, Qayyum, Syed Amir, Aaron, Mohd Adli, Noor Bahiyah and Aeslina are located below 1.81 logit of person measure and they somehow are a little bit lacking in the performance of the given test. Table V shows the summary statistics of measured persons. According to [16], the representativeness of the items can be investigate by checking the value given for item strata. It labeled as "SEPERATION" which the minimum value for item strata is 2. The separation value given for this test is 1.25 which is not an acceptable index. According to [34], sample sizes as small as 100 are often adequate for estimating stable Rasch-model parameters.

TABLE V
SUMMARY OF 19 MEASURED PERSONS

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR
MEAN	17.9	24	1.66	0.63
S.D	3.4	0	1.07	0.17
MAX.	22.0	24	3.47	1.05
MIN.	7.0	24	-1.18	0.47
REAL RMSE .67		SEPARATION 1.25		PERSON RELIABILITY 0.61
MODEL RMSE .65		SEPARATION 1.30		PERSON RELIABILITY 0.63
S.E. OF PERSON MEAN = .25				

However, 60 students were used for English vocabulary test and resulted on acceptable for the representativeness of the test items [16]. As stated by [29], IRT is much more powerful and extremely important in large-scale testing (hundreds sample sizes or larger.), but completely not for classroom-sized samples.

VI. CONCLUSION AND FUTURE WORKS

By using the IRT evaluation in tests, it was found that finding the most appropriate problem levels to each student is not a big issue. Students' ability can be assessed accurately without emphasized the level of the test items. However, according to [24], learners who experience more anxiety will affect a heavier cognitive load and receive lower test scores. Thus, instructors are encouraged to provide a supportive learning environment to enhance learning effectiveness.

The proposed learning system based on IRT provides benefits in providing learning paths that can be adapted to various levels of item difficulty and various learners' abilities. Multiple-choices test were used together with learning guidelines that provided in the system can prevent learners becoming lost in the course materials. Thus, cognitive loading also can be reduced by filtering unsuitable course materials.

Furthermore, the results help in identifying items which need for modification before future administrations. Using Rasch analysis, useable information is provided which allowed not only for separating students by ability but also contributed to the future processes of development, modification and monitoring of pedagogical assessment.

This was only a preliminary study on IRT using Ministep software and cognitive load. In the future, a larger sample size with suitable IRTs' software and wider range of subjects are recommended to validate the current findings.

ACKNOWLEDGMENT

We would like to extend sincere appreciation to TATI University College and also for all members of IT Education Research Group at Computer Science Department, University Malaysia Terengganu, for the supports and encouragement.

REFERENCES

[1] B. Baker, & Frank, *Item response theory: Parameter estimation techniques. New York: Marcel Dekker*. 1992.
[2] C. K. Hulin, C. L., Drasgow, F., & Parsons, *Item response theory: Application to psychological measurement. Homewood, IL: Dow Jones Irwin*. 1983.
[3] C. Chen, H. Lee, and Y. Chen, "Personalized e-learning system using Item Response Theory," *Comput. Educ.*, 2005.
[4] Y. Yung-Chin, H. Rong-Guey, C. Li-Ju, C. Kun-Yi, and C. Yan-Lin, "Development and Evaluation of a Confidence-Weighting Computerized Adaptive Testing," *Educ. Technol. Soc.*, vol. 13, no. 3, pp. 163–176, 2010.
[5] "Using Rasch Analysis to identify uncharacteristic responses to undergraduate assessments," *Teach. Math. Appl.*, 2010.
[6] A. Baylari and G. Montazer, "Design a personalized e-learning system based on item response theory and artificial neural network approach," *Expert Syst. Appl.*, 2009.
[7] M. Thomas, "The value of item response theory in clinical assessment: A review," *Assessment*, 2011.
[8] F.-H. Wang, "Application of componential IRT model for diagnostic test in a standard conformant elearning system," in *In Sixth international conference on advanced learning technologies (ICALT'06)*, 2006.
[9] B. B. Reeve and P. Fayers, *Applying item response theory modelling for evaluating questionnaire item and scale properties*. 2005, pp. 55–73.
[10] L. Ding and R. Beichner, "Approaches to data analysis of multiple-choice questions," 2009.
[11] G. FISCHER and I. MOLENAAR, *Rasch models: foundations, recent developments and applications*. 1995, p. New York: Springer–Verlag.
[12] J. Funk and R. Rogge, "Testing the ruler with item response theory: increasing precision of measurement for relationship satisfaction with the Couples Satisfaction Index.," *J. Fam. Psychol.*, 2007.
[13] M. O. Edelen and B. B. Reeve, "Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement.," *Qual. Life Res.*, vol. 16, pp. 5–18, Jan. 2007.
[14] J. Runnels, "Evaluation of a Streaming Instrument," *J. Kanda Univ. Int. Stud.*, vol. 25, 2013.
[15] D. Beglar, "A Rasch-based validation of the Vocabulary Size Test," *Lang. Test.*, vol. 27, no. 1, pp. 101–118, Oct. 2009.
[16] P. Baghaei and N. Amrahi, "Validation of a multiple choice English vocabulary test with the Rasch model," *J. Lang. Teach. Res.*, vol. 2, no. 5, pp. 1052–1060, 2011.
[17] R. Hambleton and R. Jones, "Comparison of classical test theory and item response theory and their applications to test development," *Educ. Meas. issues …*, 1993.
[18] A. Manuscript, "NIH Public Access," vol. 38, 2007.
[19] V. Gothwal and T. Wright, "Rasch analysis of the quality of life and vision function questionnaire," *Optom. Vis. …*, 2009.
[20] D. Kinney, "Examining Construct Stability Across Career Stage Cohorts," 2011.
[21] M. Reckase, *Multidimensional item response theory*. 2009.
[22] J. M. Linacre, *Winsteps Help for Rasch Analysis*. 2012.
[23] R. Pishghadam and H. S. Ahmadi, "Development and Validation of an English Language Teacher Competency Test Using Item Response Theory," *Int. J. …*, 2011.
[24] I. Chen and C. Chang, "Cognitive load theory: An empirical study of anxiety and task performance in language learning," *... Res. …*, 2009.
[25] F. Amadieu, C. Mariné, and C. Laimay, "The attention-guiding effect and cognitive load in the comprehension of animations," *Comput. Human Behav.*, 2011.
[26] E. Cooper, "Overloading on Slides: Cognitive Load Theory and Microsoft's Slide Program PowerPoint," *AACE J.*, 2009.
[27] F. Kirschner, L. Kester, and G. Corbalan, "Cognitive load theory and multimedia learning, task characteristics, and learning engagement: The current state of the art," 2010.
[28] B. B. Koning, H. K. Tabbers, R. M. J. P. Rikers, and F. Paas, "Towards a Framework for Attention Cueing in Instructional Animations: Guidelines for Research and Design," *Educ. Psychol. Rev.*, vol. 21, no. 2, pp. 113–140, Apr. 2009.
[29] N. A. Thompson, "Classical Item and Test Analysis with CITAS White Paper," 2009.
[30] H. Kurum, "Application of the Rasch Rating Scale Model with Mathematics Anxiety Rating Scale-Short Version (MARS-SV)," 2012.
[31] M. Sulaiman, Z. Haji Ismail, A. Abdul Aziz, and A. Zaharim, "Lesson Study : Assessing Pre-service Teacher ' s Performance of Teaching Chemistry," pp. 208–213, 2011.
[32] T. Sakumura and H. Hirose, "Test evaluation system via the Web using the item response theory," vol. 13, no. 3A, pp. 647–656, 2010.
[33] J. Baumert, M. Kunter, and W. Blum, "Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress," *Am. Educ. …*, 2010.
[34] J. M. Linacre, "Sample size and item calibration stability," *Rasch Meas. Trans.*, vol. 7, no. 4, p. 328, 1994.

A. Yacob is a lecturer and a researcher at Terengganu Advanced Technical Institute University College (TATiUC). She holds a Master of Science (Information Technology – Manufacturing) and a Bachelor of Science (Computer) at University Technology Malaysia (UTM). Her research interests include Computer programming, Quality control, education and computer industry. Her main research concentrates on Knowledge Management system.

Noraida Haji Ali received her Bachelor, Master degree and PhD in Computer Science fromUniversiti Kebangsaan Malaysia in 1995, 1999 and 2012 respectively. She is currently a lecturer in School of informatics and applied mathematics, Universiti Malaysia Terengganu. Her current research interests focus on the software engineering area especially in the systems development, Decision-Support System and object-oriented modeling include formal method. She is actively conducting her research in her field of interests through the supervision of undergraduate and postgraduate students. She has received research grants from MOHE (Ministry of Higher Education) under ERGS grant, FRGS, RACE and NRGS as a leader and co-researcher. She already completed 2 projects under university grant and 5 projects under MOHE grant as a project leader. Now, she actively applies more grants from for this year.

M. H. Yusoff is a lecturer and researcher at Department of Computer Science, UMT. He earned a Ph.D. in Computing Science from the Newcastle University, UK in 2011. He has worked for several years in the areas of e-learning system development and currently working on knowledge management system.

M. Y. MohdSaman is a Professor at the Department of Computer Science, Universiti Malaysia Terengganu (UMT). He was the Deputy Dean (Academics), between 2002 and 2006 at the same university. He received his MSc in Computer Science at UniversitiTeknologi Malaysia, and Ph.D in Computer Science (Parallel & Distribution Processing) at Loughborough University of Technology, Loughborough, England. His main research areas are in the fields of parallel & distributed computing (MPI) and network performance modelling. He has published more than 90 national and international journal papers and proceedings.

W. M. Amir Fazamin W. Hamzah currently is a Phd student at Universiti Malaysia Terengganu. He received his MSc in Computer Science at Universiti Malaysia Terengganu in 2010. His main research areas are in the fields of e-learning system development and computer human behavior.