Rasch Model Measurements as Tools in Assessment for Learning

B. Sumintono
University Malaya, Kuala Lumpur, Malaysia

Abstract—**Assessment for learning means to provide good information for teachers to help students learning better. Besides tools from the Classical Test Theory (CTT) approach that usually used by teacher, other approach which is called as objective measurement that based on probability is an alternative that can give more precise measurement. Rasch model providing psychometrics analysis techniques can be used by teachers to develop test items as well as an essential tool that can serve relevant information regard to student assessment for learning.**

Keywords— ***Model Measurements, Assesment for learning***

I. INTRODUCTION

Assessment in education is an integral process of educational activities. The learning process that occurs in schools always involves the assessment of learning as an essential thing to do. Without that, it is difficult to know for sure, whether the progress of learning has been achieved or not. Almost all the tests conducted in many schools generally use a score approach to explain student achievement. At the same time there is an unavoidable weakness with this approach that usually cannot support an effective feedback to students. The Rasch modeling measurement approach can be used, in order to provide a different perspective to the same data. This paper explains the scope of the assessment, especially the formative test and how to use Rasch model psychometrics techniques in accordance with the assessment for learning perspective.

A. Educational Assessment

The definition of an educational assessment is very diverse, but it usually mentions that decision to put the learner in a context that can state what he/she knows and is capable of (also explains what he/she does not know and has not been able to do). The definition of an educational assessment like this is so broad that it indicates that to know the progress of one's learning, it can be done both formally and informally, at any time and within a timeframe which should not be restricted [8].

The most widely recognized form of educational assessment is the test or examination. The test is a usual evaluation procedure performed by a teacher on the knowledge and skills of the students to know their performance by using certain instruments. The type of test most commonly used by teachers to their students in the classroom is a written test. There are two types of test widely known namely formative and summative assessment. Formative assessment is an assessment activity given by teachers to students where the goal is more to provide useful information to improve next learning activities. This implies that the formative assessment of teachers collects information and interprets evidence of existing learning outcomes, what students need to know more about, and adapts the teaching according to the needs of the students. In this popular language, it is also referred to a learning assessment. Meanwhile, summative appraisal is an assessment done to find out what a student already knows or what he can do, at the end of the study period. The goal is to provide information, what achievements have been achieved; in popular terms is called an assessment of learning.

The results of the test performed by students are usually used in various ways. The score of a student get in a test can show how well he or she is performing in the class, or comparison of the achievements he or she has previously. Moreover, the results of these exams can be used by teachers to: (a) determine students' abilities relative to other students in the same test; (b) showing the development of a student's ability over a period of time in certain knowledge and skills; (c) show evidence of understanding of a particular subject matter, knowledge or idea; and (d) it can predict student performance in the future. In order for the test results to be reliable and appropriate to use, then the validity and reliability aspects of the instrument are essential to be known and report.

B. Analysis of Test Results

The results of the examination analysis starts from obtaining information about students' abilities from the results of the tests conducted, usually called as ‘test score’. There are various ways to report scores that show students ability. A common way to do is to sum up the number of correct answers, which indicates students' ability. Further analysis is by performing simple statistical procedures to be able to explain more about the quality of the questions, the quality of the students as well as the comparison of attributes measured.

The most widely used approach currently in the analysis of exam results is the classical test theory (or CTT) approach. Classic test theory can be used to predict the outcome of a test. This prediction is done by considering several parameters such as students' ability and item difficulty level. Charles Spearman put forward the theory of this classic test in 1904 and applied in many discipline including educational assessment. The basic assumption of this classical test theory is that the observed scores are denoted by X, none other than true scores (T) and errors (E), so the equation: X = T + E [1].

This means that score of test results obtained by one student, for example, contained true scores and errors. It should be noted that only the observed score (X) is real (appears in the data directly) while the true score (T) and error (E) are latent or cannot be observed directly. From these observed score (which is a raw score), various analysis and interpretations can be produced such as: a) descriptive statistics, i.e. central tendencies (average), variance and frequency tables. All three can provide information directly on which items are useful and which are not. For example, the low diversity of scores among students indicates poor quality of item questions in the test; b) item difficulty level; the degree of item difficulty shows the proportion of students who can answer the item correctly from one exam. The lowest point of 1.0 (100%), meaning that all students can answer correctly about the test and the highest point of difficulty level is 0.0, indicating none (0%) individuals who can answer correctly. Item difficulty that have an extreme point (0% or 100%) like the two preceding examples are of little use because they cannot distinguish individual abilities, in other words they are not good quality items; c) The item discrimination shows how far a problem is able to distinguish individuals with high and low ability. Simply put, if high-ability and low-ability students can overcome item number 10, then this problem has low item discrimination. On the contrary, if a high-ability student can solve the item problem number 10 while the low-ability cannot cope, then point the item has a high discrimination; d) weighted score, generally in the context of CTT, the scores for each item are given equally (e.g. 1 for correct answers, and 0 for wrong answers), weighting scores are applied when a given problem has different weights to produce a total raw score.

Basically, the use of raw score as a measure of achievement has several disadvantages, such as [1]:

a. The raw score is basically not the result of measurement. More precisely the raw score is the number of correct answers to the item questions;
b. The raw score is the initial information. The raw score is also usually expressed in percentage (%) which is nothing but a summary of numerical data, but does not provide measurement data;
c. Raw scores have weak quantitative meanings. The quantitative meaning of the raw score obtained will be different, depending on the number of questions, while the percentage of correct answers always depends on the difficulty level of the problem;
d. The raw score does not indicate a person's ability to a particular task. The raw scores also can not explain much about the difficulty of the problem; and last,
e. The raw score and the percentage of correct answers are not always linear. In a linear test, students who score 15 (scale 0 to 100) always have a higher ability than those with a score of 10. However, empirically sometimes both have the ability to have the same.

More critics come from van Zile-Tamsen (2017, p. 2), she stated that CTT approach has several limitations:

"*including the fact that derived scores are sample dependent and biased toward central scores. Further, missing data presents a problem for computing overall scores. Measure reliability is often presented as Cronbach's alpha, and evidence of validity is based on the content of the items and correlations of scale scores with other measures, which may or may not be reliable and valid themselves. Finally, it is very difficult to examine the operation of individual items to determine effectiveness of these items for the target population and their contribution to measurement of the overall latent construct.*"

Therefore looking at other alternatives in conducting analysis of exam results is indispensable, especially with the various weaknesses of the classic test theory above. The deficiency of CTT is then corrected by the theory of item response theory (IRT) with various variations of its logistic parameters (called as PL), one of which is 1PL developed by Georg Rasch that called as a rasch model. Unlike the CTT which always depends on the score, IRT is not dependent on the sample of particular item-questions and abilities of people involved in the exam.

II. METHOD

Georg Rasch developed an analytical model of the item response theory (IRT) in the 1960s commonly called 1PL (one parameter logistic) [9]. This mathematical model was later popularized by Benjamin Wright in the United States (Linacre, 2011). With raw data in the form of dichotomous data (in the form of right and wrong) that indicate the student's abilities, Rasch formulates this into a mathematical model that connects students and item interchangeably trough an equal interval scale [11].

As an illustration, a student who is able to do 80% of the problem correctly would have better abilities than other students who can only do 60% of the item questions. The data gathered (percentage) indicates that the raw data obtained is none other than the ordinal data type showing rank and not linear [5]. Since ordinal data do not have the same interval, the data needs to be converted into equal interval scale for statistical analysis purposes. So if a person gets a score of 80%, then the odds probability ratio is 80:20 (meaning: 80 correct and 20 wrong), which is nothing more than a more an odd ratio probability. However, this odd ratio score still not has equal interval characteristics, so Rasch suggests to use logarithmic function to produce measurements scale with the same interval (equal distance). The result is an equal interval scale that also has a new unit called logit (log odds unit). Through this interval type data, Rasch model develops a measurement model like a logit ruler that determines the relationship between the student ability and item difficulty level. In practice this interval data showing students' abilities and item difficulty in the same scale. Later, based on this model it is easy to be concluded that the success rate of students in working on the test items depends on the level of ability and item difficulty level [4].

The results of Rasch measurement model through logit ruler addresses the five principles of measurement for human sciences from Mok dan Wright (2004), which are: a). produce a linear measure; b). overcome missing data; c). give estimate of precission; d) detect misfits or outliers; and e). replicable [7]. If the examination analysis which starts from obtaining information about students' abilities that follow this principle, meaning more accurate and meaningful inferences can be made on the data that gathered, especially for activities of assessment for learning.

III. RESULTS AND DISCUSSION

A. Wright Map (Item-Person Map)

Item person map (or Wright Map or Variable Map) is a tool in Rasch model measurement that provide comprehensive outlook of the data. This map, also called as construct map, illustrates person abilities and item difficulties which using the same logit ruler that provide information about result of a test [13].

For illustration, theoretically, the continuum example of the item difficulty level can follow what in education called as Bloom's Taxonomy. In the 1950s Benjamin Bloom proposed a taxonomy of cognitive process. This taxonomy is so influential in education, and has undergone various revisions. According to Bloom, the items that ask about memorizing categorize as the lowest level of cognitive ability. Therefore the items that measure this process tend to have low difficulty levels. The higher the level of cognitive processes performed, the higher the degree of difficulty of the item questions that measure it. The level of cognitive processes developed by Bloom moves from memory, understanding, application, analysis, evaluation and finally synthesis. This means that the test item synthesis type should be the most difficult to be done properly by students.

Look at the Figure 1 below, that illustrate about person ability relate to item difficulty in the context of cognitive process. The left side is person ability, and the right side of the map is item difficulty level. For the person with average cognitive ability, it tends can solve correctly items that in bloom taxonomy is items type of memorizing, understanding and aplication. Meanwhile for the person who have low cognitive ability (left side of map in the bottom), the person has high probability only to solve correctly item question relate to memorizing facts. This map can easily capture the whole picture about person ability and item difficulty situation in one occasion.

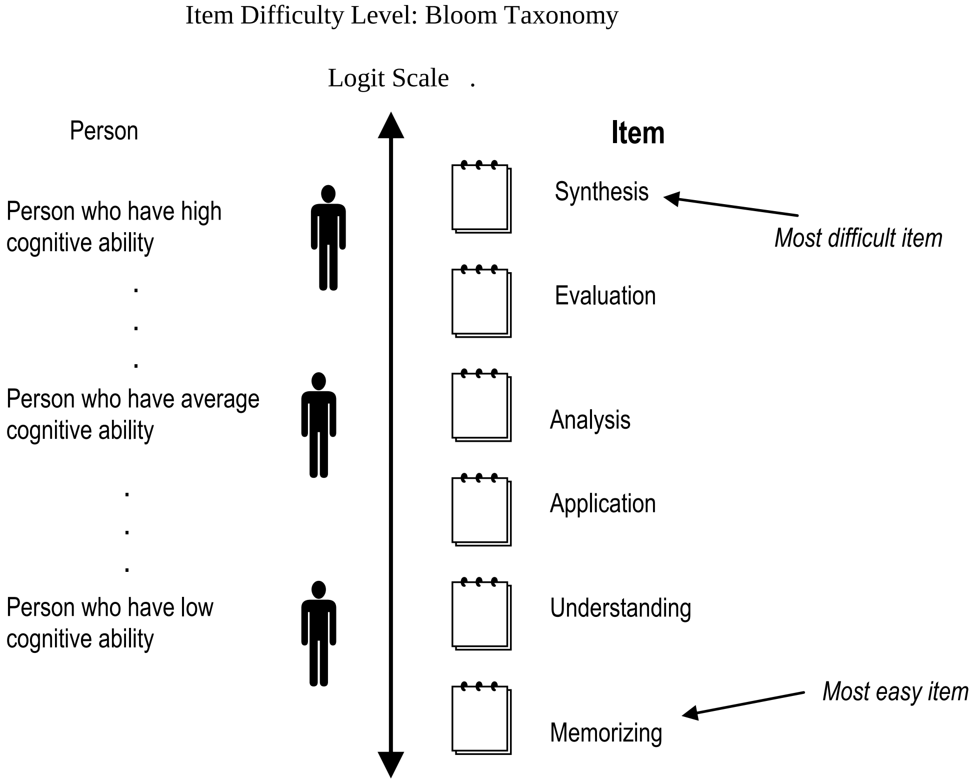
Item Difficulty Level: Bloom Taxonomy
Logit Scale .
Person
Item
Person who have high cognitive ability
Person who have average cognitive ability
Person who have low cognitive ability
Synthesis
Most difficult item
Evaluation
Analysis
Application
Understanding
Memorizing
Most easy item

Fig 1. Bloom Taxonomy Construct Map

B. Instrument Development

Rasch measurement model is an alternative to the development of measurement instruments on educational assessments other than using classical theory. Some of the steps typically passed in the measurement instrument development procedure are:

1) Verify the assumption of unidimensionality and local independence of measurement

2) Testing the accuracy of individual item in the model. Item that have a low accuracy value are removed from the analysis (having bigger standard error measurement). The analysis is repeated again until all items have good precision with the model.

3) If the remaining number of item still exceeds the number of items being targeted, we may select the items by various considerations, for example: (a) items not overlapping their location with other items (have the same item difficulty level), (b) items that can improve the measurement reliability, - the response according to the sequence (to examine the graph of the item characteristics) or (d) items that provide information in accordance with the measurement function.

The evaluation process of the measurement instrument is an iterative analysis process, which is done repeatedly until the researcher finds the optimal composition, where all the criteria can be fulfilled. For-instance, a good instrument is having items that contain from lowest difficulty to the highest difficulty level; then with this will measure precisely person ability that come from every level of ability spectrum.

In practical terms, according to Boone, Staver and Yale (2014), the criteria used to check the suitability of item that could be outliers or misfits is refer to three psychometric attribute of each item that can be generated from Racsh model software such as Winsteps. The three psychometrics attribute are an Outfit mean square value (0.5 <Outfit MNSQ <1.5), a Z-standard Outfit value (-2.0 <ZSTD <+2.0) and a Point Measure Correlation Value (0.4 <Pt Measure Corr <0.85). Item that has the value beyond of these three psychometrics attribute, can be categorize as misfit items that need to be revised and re-test again.

C. Detecting Item Bias

Items and measurement instruments can be biased, i.e. when an item is more favorable to one group of certain characteristic than the others. A test item that explains about making batik, will be easy to understand by student who come from Java compare to other parts in Indonesia. This means the item is bi-as because it easy to answer by Javanese students than other ethnicities. This item tends to be biased in measuring, which in psychometrics is called the item has a differential item functioning (DIF). Rasch modeling provides a tool that can detect the presence of bias (DIF) based on the response given to certain items based on demographic data of respondent provided.

In the Winsteps software for instance, many demographic data can be combined to detect item bias, for example gender with domicile, which will give very good information based on this characteristics in terms of students' ability in this groups. Practically an item called has DIF (bias) when value of its DIF-probability less than 5% (0.05). At the same time, because DIF gives information about item difficulty level for each item based on demo-graphic profile of respondent, this will be a very handy analysis to map overall ability based on stu-dents characteristics.

D. Person Diagnostic Report

In addition to measuring item difficulty, Rasch model also can measure ability of individual more precisely. The accuracy from response given, the pattern will show about individual tendency regard to how he/she perform solving test items. In this aspect a teacher can find out information from the results of the tests performed, where the formative test will provide valuable information for improving teaching and helping students more precisely. Detection of test result can be done in the form of identification of misconception of students on certain subject, which can be known from the statistical fit information (psychometric attributes) and the pattern of responses that are out of the ordinary.

Tools that can be used for this is called scalogram, that systematically present result of each individual responded to each item in the test. Pattern that shows up in the scalogram could be person who is consistent in their cognitive ability (solving problem correctly from item with low difficulty to high difficulty level). In other situation, could be identify person who have misfit characteristics such as wrongly answer low difficulty level item (careless situation), or can solve difficult item by the person who has low ability (an indication of lucky guess).

Another tool that inform comprehensively for each individual item is person diagnostics table (Figure 2 above). This individual report showing person ability (look at the Figure 2), which denote as XXXX in the middle where the ability is +0.57 logit. The horizontal line above and below XXXX is the standard measurement error of this person (0.80), which show his/her highest and lowest ability from the measurement value of +0.57 logit. Item number in the left side are items that answer correctly (item number 4, 9, 8, 5, 1 and 10), whereas items on the right side are answer not-correct (6, 7, 2 and 3). If the item number position in the bottom (such as no 4 and 9), then it is easy item; but if the position at the top area of the table (item no 3) then it is a difficult item.

The person diagnostics report showing that based on this student ability, item no 6 and 7 should be done correctly, because its difficulties are below his ability. For no 2 and 3 it is make sense if wrong, because item difficulty level is higher compare to person ability.

IV. CONCLUSION

Instrument testing and determining students' abilities in educational assessments are essential. An analysis that can result in more precise measurements (produce an equal-interval scale) will determine the quality of the results of the analysis and the improvement of the educational process to help students learning. The Rasch model can help teachers assess in improving the quality of the analysis performed, because it applies the appropriate basic principles data processing. This is because the Rasch model addresses to five objective measurement requirements.

Rasch modeling applications in formative test have many advantages as they seriously about measurement accuracy. This can be for the detection of item difficulty and item bias, as well as on individual abilities identification and provide learning assistance appropriately.

REFERENCES

[1] S. Alagumalai, D.D. Curtis, and N. Hungi, (editors*) Applied Rasch Measurement: book of exemplars*, papers in honour of John P. Keeves, Dordrecth: Springer, 2005.

[2] T.G. Bond, & C. Fox, *Applying the Rasch Model. Fundamental measurement in the Human Sciences*, 3rd edition, New York: Routledge, 2015.

[3] W. J. Boone, J.R. Staver and M.S. Yale, *Rasch Analysis in the Human Sciences*, Dordrecht: Springer, 2014.

[4] G. Englehard, *Invariant Measurement, using rasch models in the social, behavioral and health sciences*, New York: Routledge, 2013.

[5] J. M. Linacre, "Investigating rating scale category utility," *Journal of Outcome Measurement,* vol. 3, no.2, pp. 103-122, 1999.

[6] J.M. Linacre, *A User's guide to WINSTEPS Ministeps*; Rasch-model Computer Program, Program Manual 3.73, 2011.

[7] M. Mok, and B. Wright, Overview of Rasch Model Families. In *Introduction to Rasch Measurement: Theory, Models and Applications* (pp. 1-24). Minnesota: Jam Press, 2004.

[8] D. Musial, G. Nieminen, J. Thomas, dan K. Burke, *Foundations of Meaningful Education Assessment,* Boston: McGraw-Hill Higher Education, 2009.

[9] L. W. Olsen. Essays on Georg Rasch and his contributions to statistics. Unpublished PhD thesis at Institute Of Economics University of Copenhagen. 2003.

[10] B. Sumintono, dan W. Widhiarso, *Aplikasi Model Rasch untuk Penelitian Ilmu-ilmu Sosial (edisi revisi),* Cimahi: Trim Komunikata Publishing House, 2014.

[11] B. Sumintono, dan W. Widhiarso, *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*, Cimahi: Trim Komunikata Publishing House, 2015.

[12] C.V. Zile-Tamsen, Using Rasch Analysis to Inform Rating Scale Development, *Research in Higher Education (pp 1-12). Published online 20 Feb 2017,* DOI 10.1007/s11162-017-9448-0, 2017.

[13] M. Wilson, *Constructing Measures, an item response modeling approach*, New Jersey: Lawrence Erlbaum Associates, 2005.