English Language Assessment and Rasch Model Measurement: an Introduction

Bambang Sumintono
University Malaya
bambang@um.edu.my

Introduction

In the modern and globally interconnected world, use of *lingua franca* is inevitable. Exchange of goods, money, ideas and expertise is happening every minute trough digital connection where English is the main language. This resulted to the big effort to maintain standard of language proficiency where English language testing so wide speared across the globe. Pricillia Allen (2009) writes that, "language testing is the practice and study of evaluating the proficiency of an individual in using a particular language effectively," which shows complexity of assessing person ability to speak in other language. This position papers try to introduce of a measurement model that happening in English language testing (McNamara & Knoch, 2012), that not only resulted to estimate individual language proficiency, but also involve testing fairness even relate to social justice.

English Language Assessment

In the context of language ability, McNamara (1996) offers perspective that all language models have three dimensions, namely knowledge, ability for use or performance, and actual language use. Many authors (for example Fulcher & Davidson, 2007; Bagarić & Djigunović, 2007) grouped the first two dimensions as communicative competence or communicative language ability. As defined by Hymes (1972) communicative competence is the ability to use rules of grammar not only accurately but also appropriately according to the communicative events. Further, Canale and Swain (1980) that develop communicative competence ideas that proposed the first model of communicative

competence which should be consisted of three components: (a) grammatical competence; (b) sociolinguistic competence; and (c) strategic competence.

In the situation of language assessment and testing grammatical competence is measured by the ability of test taker (testee) to recognize and manipulate lexical items and morphological, syntactical, semantical, and phonological rules which challenge them to understand rule of the game other language. The second component, sociolinguistic competence is concerned with appropriateness of language use in terms of meanings and forms within specific social contexts, where this situation need social and cultural situation of other language user which probably beyond test taker experience and imagination. Strategic competence, the last component, involve not only verbal but also nonverbal strategies that need to be used to make it message more effective to deliver such as paraphrase (verbal) or gestures (non-verbal) (Canale & Swain, 1981).

In more practical form, person ability in other language proficiency is mostly measure in the four macro language skills, which are reading, listening, speaking, and writing. To make good estimation of test-takers' other language ability, their responses to the test items that consist of those macro language skills is measured. Type of test items for measuring language competence in reading and listening most commonly used multiple choice format or true/false statement, which number of correct answer showing the ability in general. However, regard to speaking and writing ability which showing a kind of product of language competence, the same type of test cannot be applied similarly. Different type of assessment which rely on rubric and rating seem more appropriate to be used (Finch & French, 2019; Engelhard & Wind, 2018). Something emerged in the last forty years in educational assessment such as in English language assessment field, was the need to apply more precise measurement model (see Leonard, 1980), which the culmination of this measurement model as the new standard (McNamara & Knoch, 2012).

Rasch Measurement Model

Georg Rasch developed an analytical model of item response theory (IRT) in the 1960s which later called as Rasch Model which is a variation of IRT with 1PL (one logistic parameter) model (Olsen, 2003). This mathematical model was later popularized by Ben Wright in the United States of America. With raw data in the form of dichotomous data (in the form of right and wrong) that indicate the ability of students, Rasch formulates this into a model that connects students and items (Sumintono & Widhiarso, 2014; 2015).

As an illustration, a student who is able to do 80% of the questions correctly, certainly has better ability than other students who can only answer 60% of the questions. The data (percentage) shows that the raw data obtained is none other than ordinal data types that show rank and are not linear (Linacre, 1999). Because ordinal data does not have the same interval, the data needs to be converted into ratio data (probabilistic data) for statistical analysis purposes. So if someone gets an 80% score, then the odds ratio is 80:20, which is none other than the ratio data of right answer divide by wrong answer, that is more appropriate for measurement purposes. Through this ratio data, Georg Rasch develops a measurement model that determines the relationship between student ability level and item difficulty level by using the logarithm function to produce measurements with the same interval. The result is a new unit called log odds unit (log odds unit) which shows student ability and item difficulties using the same scale (a logit scale); so that later from the logit value obtained, it is concluded that the level of success of students in working on the problem depends on the level of ability and the level of difficulty of the problem (Olsen, 2003).

For data in the form of a dichotomy, Rasch modeling combines an algorithm that states the results of probabilistic expectations from the 'i' item

and the 'n' respondent, which is mathematically expressed as (Bond & Fox, 2015)

$$P_{ni}(x_{ni}=1/\beta_n, \delta_i) = \frac{e^{(\beta n - \delta i)}}{1 + e^{(\beta n - \delta i)}}$$

In social sciences research, data can be obtained through a cognitive test instrument such as ability and intelegence. Cognitive test instrument mainly regard a test either in the form of low stakes (diagnostic test) or high stakes (public examination) that usually in the form of dichotomous such as True and False or multiple choices; to measure ability more precise test maker also develop rubric that like rating scale. The instrument is designed to measure a variable that have been defined satisfactorily, then identified its relevant constructs; from there, items are created, tested and developed to be able to measure spectrum of the variable. The answer choices provided generally follow the scoring pattern adopted by the classical test theory (CTT). In the context of the rasch model, this scoring pattern treated as raw data in ordinal type, where each item and person processed to find its odd probability, then transforming into logit using logarithm function. The product of this process is person measure (person logit) and item measure (item logit), which following a measurement model what called as objective measurement in quantitative research in social science.

The logit scale generated in the Rasch model is a scale with the equal-interval and is linear derived from the data ratio (odds ratio) and not the raw data obtained (1). Therefore, the process of estimating one's ability or level of difficulty will have a more precise estimation value and can be compared to each other because it has the same unit (logit) (2). Since the algorithm used will sort structurally between respondents from high to low ability, which simultaneously also sort the item from easy to difficult, then if

there is an inaccuracy/consistency of answers from the respondent (misfit) or out-of-pattern (outlier) easy to detect (3). The order of respondent's ability and structured problem difficulties also make the rasch model predict when there is missing data (4). The resulting logit scale will bring up a value that depends on the response pattern provided, rather than on the initial score specified, so that the rasch model will always produce independent measurements (5).

Further, the above description about 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. 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. Because of this the quality of measurement in social science carried out with a Rasch model will have the same quality as the measurements made in the field of physics.

Analysis with the Rasch model produces a fit statistics analysis that provides information to the researcher whether the data obtained does ideally illustrate that people who have high ability provide patterns of answers to items according to their level of difficulty. The parameters used are infit and outfit of the mean square and standardized values. According to Sumintono and Widhiarso (2014), infit (inlier sensitive or information weighted fit) is the sensitivity of the response pattern to the target item on the respondent (person) or vice versa; while the outfit (outlier sensitive fit) measures the sensitivity of the response pattern to items with a certain level of difficulty in the respondent or vice versa.

Quantitative research in social science always faces fundamental criticism in terms of testing its research instruments. The quantitative test instrument commonly used in CTT is the reliability index (Cronbach's alpha)

which only measures the interaction between items and persons; how good quality of individual item can never be done because there is no measurement index that can used at that level; also at the same time to detect inconsistent respondent answers is not available. It is different from the classical test theory, in rasch the item analysis model is carried out to the level of each item. In addition to items, the Rasch model also simultaneously tests the person (respondent), where the respondent's pattern of responses is it consistent or not (Bond & Fox, 2015). Tests for research instruments can also be carried out in the form of dimensionality tests, the rating scale analysis or the detection of bias from the items tested. All of this can be done because basically the Rasch model fulfills all objective measurement requirements.

Rasch Model Application for Instrument Development in Language Assessment/testing

1. 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/agreeability and item difficulties which using the same logit ruler that provide information about result of a test (Wilson, 2005).

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 application. 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.

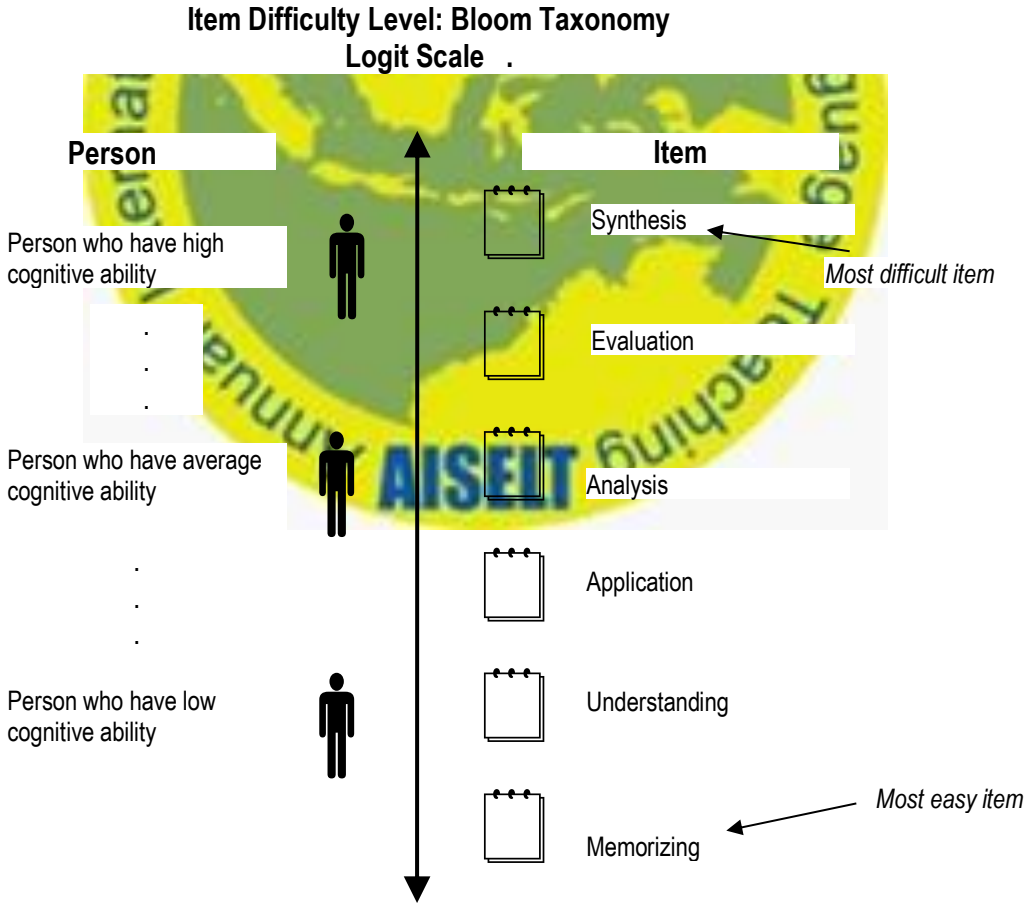

Figure 1. Bloom Taxonomy Construct Map

What also important in Figure 1 is that a good instrument has the capacity to measure all probability of abilities of people (low-middle-high ability). This means, an instrument with so many variance of items difficulty has the good quality to measure people's ability. In rasch model this called as a good construct validity (Linacre, 2017).

2. Item Fit statistics

Rasch modeling is a good alternative for developing instruments for cognitive test compare to classical test theory. Some of the steps that are usually passed in the procedure for developing measurement instruments are:

a) Verification of assumptions about local uni-dimensionality and independence of measurement

b) Testing the individual item accuracy with the model. Items that not fit with certain range of value has low precision are excluded from the analysis (quality control). The analysis is repeated again with different set of data until all items have accuracy with the model.

c) If the remaining number of items still exceeds the number of items targeted, then we can select items with various considerations, for example: (a) items that do not overlap their location with other items (has the same item difficulty level), (b) items that can improve measurement reliability, items that are options - response responses are in the order; or (d) items that provide information that matches the measurement function (analyzing the test information function graph).

The evaluation process of measurement instruments is an iterative analysis process, which is carried out repeatedly until the researcher finds an optimal composition, where all criteria can be met. In Winsteps software program, unidimensionality is found in Item function: dimensionality (Table 23) and accuracy of items with model (infit-outfit) and location (measure)

can be seen in Item: measure (Table 13) or Item: fit order (Table 10) (Linacre, 2011).

3. Measurement Bias Detection

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 process of making *batik*, will be easy to understand by student who come from Java compare to other parts in Indonesia. This means the item is bias 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 demographic profile of respondent, this will be a very handy analysis to map overall ability based on students characteristics (Linacre, 2011).

4. Multi-rater analysis

Another challenge for researcher is when their research design involve in using multiple judges or raters. Traditionally analytical tools used to analyze this multi rater data are using Kappa Cohen/Kappa groups and intra-class correlation coefficient. However, these analyses cannot detect if we want to know about raters severity and leniency precisely; also try to find misfit item used in judging subject or to know how better quality of a subject compare to others cannot be obtained (Englehard, 2013). This reflect that in multi rater situation, fairness and justice is assessment is central

issue that need to be handle carefully. Extensive study by Scullen, Mount and Goff (2000), found out that using classical test theory, 2350 managers who being assessed by 7 raters, only found maximum information around 62% about raters, information about rates and items less from that.

For multi rater analysis rasch model offer Multi Facet Rasch Model (MFRM), that developed by Mike Linacre, which provided much better analysis. By the MFRM analysis, result provided can explain how severe and lenient raters in assessing the items, assessing the level of raters' consistency, correcting test participants' scores by the severity raters, assessing the functioning of rating scale, and detecting raters' bias interactions (Englehard, 2013, 2018; Bond and Fox, 2015). The result shows that fairness and justice in assessment that involve human judgment can be accommodated nicely with the MFRM.

References

Allen, P. (2009). Language testing definition, URL: http://languagetesting.info/whatis/lt.html

Bagarić, V., & Djigunović, J. M., (2007). Defining communicative competence. *Metodika, 8*(1), 94-103.

Bond, T.G., & Fox, C. (2015). *Applying the Rasch Model. Fundamental measurement in the Human Sciences (3rd edition)*. Lawrence Erlbaum Associates, Publishers. Mahwah. New Jersey

Boone, W. J., Staver, J.R., and Yale, M.S. (2014). *Rasch Analysis in the Human Sciences.* Dordrecth: Springer.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Canale, M., & Swain, M. (1981). A theoretical framework for communicative competence. In A. S. Palmer, P. J. M. Groot, & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence.* Washington, DC: Teachers of English to Speakers of Other Languages.

Englehard, G. (2013). *Invariant Measurement, using Rasch Modelsin the social, behavioural and health sciences.* New York: Routledge.

Engelhard, G.Jr and Wind, S.A. (2018). *Invariant Measurement with Raters and Rating Scales.* New York: Routledge.

Finch, W.H. and French, B.F. (2019). *Educational and Psychological Measurement.* New York: Routledge

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* Oxon, OX: Routledge.

Hymes, D. (1972). On communicative competence. In J. B. Pride, & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269-293). Harmondsworth: Penguin.

Leonard, M. (1980). Rasch Promises: a Layman's Guide to the Rasch Method of Item Analysis, *Educational Research*, vol. 22:3, pp. 188-192

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*. 3(2), 103-122.

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

Linacre, M. (2017). Teaching Rasch Measurement. *Rasch Measurement Transaction 31 (2) 1630-1631.*

McNamara, T. F. (1996). *Measuring second language performance*. London, UK: Longman

McNamara, T and Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*. Vol. 29(4), pp. 555-576.

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

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

Scullen, S.E., Mount, M.K. and Goff, M. (2000). Understanding the latent structure of Job Performance Ratings. *Journal of Applied Psychology, 85 (6), 956-970.*

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

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