Pertanika J. Soc. Sci. & Hum. 27 (3): 1469 - 1480 (2019)

PERTANIKA JOURNALS

SOCIAL SCIENCES & HUMANITIES

Journal homepage: http://www.pertanika.upm.edu.my/

Using Rasch Model to Assess Self-Assessment Speaking Skill Rubric for Non-Native Arabic Language Speakers

Rizki ParahitaAnandi* and Muhammad Azhar Zailaini

Arabic Language Education, Faculty of Education, University of Malaya, 50603 Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur, Malaysia

ABSTRACT

This study was conducted to assess the quality of the self-assessment speaking rubric adapted by Montgomery from Bill Heller in 2000. The rubric consisted of six aspects with a four-point rating scale and was originally written in English and aimed to be used by the English language learners. As the respondents of this research were the Indonesian students who learn Arabic language as a foreign language, the rubric was therefore modified and translated into Indonesian language. Rasch measurement model approach provides various analyses with empirical evidence about the quality of instrument by looking at the rating scale analysis, summary statistics, item fit, principal component analysis and Wright map. About 43 Arabic language learners from a university in Salatiga, Indonesia, were involved in this study. The finding showed that the four rating options were clearly understood by the respondents. All six items in the rubric were also appropriately measure students' speaking skills. High value of person (0.84) and item reliability (0.94) indicated good quality of both respondents and instrument. The Cronbach alpha value 0.83 indicated high reliability. To sum up, the self-assessment speaking rubric has a good quality to measure speaking skills and is appropriate to be used by students to self-assess their Arabic speaking ability.

Keywords: Arabic language, Rasch measurement model, self-assessment, speaking

ARTICLE INFO

Article history:
Received: 12 September 2017
Accepted: 19 February 2019
Published: 13 September 2019

E-mail addresses:
rizkiparahita@gmail.com (Rizki Parahita Anandi)
azhar@um.edu.my (Muhammad Azhar Zailaini)
*Corresponding author

INTRODUCTION

Many types of assessment have been

ISSN: 0128-7702
e-ISSN: 2231-8534

used to measure speaking ability. The assessment itself could be done either by the foreign language teachers or the students themselves, which is called as self-assessment. According to Blanche and Merino (1989) the first research on self-assessment (SA) was published for the first time since a long time ago in 1976, and it has continued to be used in L2 learning and education as well.

Gardner (2000) defined three different types of SA; they are teacher-prepared assessment, generic assessment and learner-prepared assessment. For the teacher-prepared assessment, all the criteria, content and instructions are prepared by the teacher, while students only need to do the assessment. Generic assessment means that the teacher stated only the criteria, while the content and assessment are done by the student. In student-prepared assessment, all work is done by students. Teachers give all autonomy to students to think about what criteria and content to be assessed; and at the end, students themselves are required to do their self-assessment.

Szyszka (2011) argued that there was a critical need to develop self-evaluation abilities among teacher trainees so that they would be able to use these abilities later when they became language teachers. Leger (2009) suggested that using self-assessment in the teaching and learning process would enable both learners and teachers to reflect on the learning process besides enabling mutual feedback. Gardner (2000) also encouraged foreign language learners to be equipped with good self-assessment skills. It might help them to monitor their progress toward specific learning objective like speaking skill. Through self-assessment, learners would discover more about the specific aspects they need to improve then asking for help from their teachers.

Teacher-Prepared Self-Assessment Speaking Skills Rubric

Several established teacher-prepared assessment rubrics could be used to assess speaking skills in the classroom. Among them is the speaking self-assessment rubric adapted by Cherice Montgomery in 2000 from Bill Heller. This rubric was intended to be used by foreign language learners to assess their own ability in speaking. The rubric was developed to measure six aspects of speaking skill, namely: 1) pronunciation (Aspect1), 2) fluency (Aspect2), 3) vocabulary and circumlocution (Aspect3), 4) accuracy and comprehensibility (Aspect4), 5) content (Aspect5), 6) comprehension and strategy competence (Aspect6).

Fluency in speaking occurres when a speaker engages in a meaningful interaction with other people and he or she is able to maintain comprehensible conversation and keep the communication ongoing in spite of her or his lack of communicative competence (Richards, 2006). Students with good fluency in speaking will be able to speak in a foreign language well in front of their teachers and peers without many pauses which may interrupt the communication process.

The next aspect is accuracy, which means the ability to speak using grammatically correct sentences. Learners should not only

know the correct grammatical rules of the language but also be able to speak accurately (Srivastava, 2014). When people intend to communicate orally, they have to master the grammatical rules so that the listeners will not misunderstand the conversation.

Pronunciation deals with how to pronounce the language while speaking (Montgomery, 2000). The ability of speaking is the combination of correct pronunciation and intonation and directly affects the appropriate communication in conversation (Zhang & Yin, 2009). Someone with good pronunciation may sound like a native speaker. On the contrary, people with poor pronunciation may lead others to misunderstand what they are saying.

Good speakers should be able to correctly use a wide range of vocabulary. Montgomery (2000), in her rubric, measured students' capability in speaking by considering the vocabulary used by the students. Students who mastered a wide range of vocabulary would be more expressive in speaking a foreign language. To assess the content of speaking, she stated that someone with good mastery of the content would be able to explain something with detailed description.

The last criteria that will be assessed in speaking are comprehension and strategic competence. This deals with how someone is able to make good conversation in any kind of situation with other people (Montgomery, 2000). Montgomery added that students with good strategic competence will be able to manage their concentration while speaking though interrupted in the middle of speaking.

Some issues regarding self-assessment still exist in the teaching and learning process. Gardner (2000) stated that the issue was about validity and reliability, because assessments were only useful if they were accepted as valid and reliable. Some researchers have found learner self-assessment to be valid and reliable. Bachman and Palmer (1989), for instance, developed an instrument for English language learners to measure their communicative language ability and proved that self-ratings were reliable and valid. On the other hand, Dickinson (1987) stated that assessment performed by teachers and other specialists was likely more reliable than those by the learners.

To prove that the self-assessment speaking rubric is appropriate to be used and the data collected from it can be used by both teachers and leaners, further analysis is required. The empirical evidence of validity and reliability analysis of an instrument will help to determine the quality of data collected from the respondents.

Validity has been defined as the development of sound evidence to demonstrate that the test interpretation matches its proposed use (Creswell, 2012), while reliabilityrefers tothe degree of consistency with which the instrument measures what it is intended to measure (Ary et al., 2006). Rasch measurement model approach provides further analysis to assessthe validity and reliability of an instrument. As Bond and Fox (2015) stated, the Rasch measurement model

helps researchers to determine the extent to which the instrument actually measures the construct or latent trait under examination. While for the reliability, Rasch measurement model will produce the Cronbach's alpha value as well as item and person reliability. The item and person strata will also be described. Item strata is the number of item or person which reflects the measurably distinct groups of items or persons from the Rasch analysis (Bond & Fox, 2015).

This study was aimed at assessing the quality of self-assessment speaking skill rubric by providing empirical evidence about its validity and reliability using the Rasch measurement model approach. The objectives of the study are as follows: 1) to investigate the function of rating scale categories of the instrument, 2) to test the reliability of rubric items using Rasch measurement model, 3) to investigate the item fit of six aspects of speaking as stated in the rubric, 4) to investigate the unidimensionality of the rubric, and 5) to identify the distribution of both person and item in the map.

MATERIALS AND METHODS

Participants

This study was carried out in order to collect quantitative data from the respondents by distributing the self-assessment speaking rubric. A total of 43 Arabic language education students were purposively chosen to participate in this study since they were attending the Arabic speaking class or also known as *muhadatsah* class at the State Islamic University in Salatiga.

Instrument

The self-assessment speaking rubric used in this research was the one from Montgomery (2000). This instrument has undergone many changes for improvement. This self-assessment speaking rubric consists of six aspects of speaking skill and each aspect has four competence levels, namely memorized, guided, responsively adapted and spontaneously improvised. The rubric was originally written in English and was aimed to be used by foreign language learners. The researcher decided to use this rubric because Arabic language had also been learnt as a foreign language in Indonesia and the aspects of speaking skill being assessed were identical. In order to meet the objective of this research, the rubric was translated into Indonesian language and some modifications were made. Two experts were selected to check the translation and decide whether or not the students could understand it. Some words were changed based on the suggestions from the experts to make it more understandable by the students.

Data Analysis

All the data collected from the respondents were analysed using the Rasch measurement model approach by using the Winstep version 3.73. Rasch model provides a set of analyses to test the validity and reliability of the self-assessment speaking rubric. The Rasch analysis was conducted in 5 stages, they are 1) rating scale analysis, 2) summary statistics, 3) item fit, 4) principal component analysis, and 5) Wright map.

RESULTS

Rating Scale Analysis

To check the instrument validity and reliability is very important before conducting the real study because the quality of the data collected would also depend on it. Anyone who develops an instrument would also consider what type of scale or response option would be the most suitable for the objective of the instrument. Some might often use a simple "yes/no" response, while others use frequency scale "never/ sometimes/often/always" and some others might consider using a Likert type scale from "strongly agree to strongly disagree". Thus, the rating scale or response option used in the instrument must be tested empirically to check whether the response options are unambiguous and whether respondents could differentiate between the options given (Bond & Fox, 2015).

Linacre (1999) had explained that the initial stage needed to be done before doing further analysis about instrument in investigating the function of rating scale categories used in the instrument. The Rasch measurement model approach provided empirical analysis of rating scale. Figure 1 illustrated the rating scale analysis of self-assessment speaking rubric that puts four rating scales in every item.

Rasch measurement model enabled us to check whether the four rating scales used in the rubric should be collapsed or separated as well as whether these four rating scales were understood by the respondents. The value of the observed average that increased monotonically from -4.46 which was negative value to 3.63 which was positive value indicated normal response from the respondents.

The value of the Andrich Threshold calibration appeared in the next column also needs to be taken into consideration. The *s* values or distance between two categories as written in the column Andrich Threshold must be in the acceptable range of $1.4 < s < 5.0$ (Aziz et al., 2013). The gap from one scale to another scale was still in the acceptable range, and it could be concluded that the four rating scales were clearly understood by the respondents. The figure above also showed that all the rating scales peak above 60% (red line) showing that respondents understood fully the different choices of rating.

Fit statistics also offered another criterion for examining the quality of the rating scale. The value of outfit mean-square bigger than 2 means that the particular category is introducing more unexpected noise than the expected noise into measurement process (Linacre, 1999). Table 1 shows that all the outfit mean square values were less than 2, indicating that the quality of four rating scales of the rubric is good.

Summary Statistics of the Rubric

Table 2 describes the summary statistic of the rubric:

The table shows that the mean was -1.07, which was below the mean of item (0.00), indicating that many students assumed their ability in speaking Arabic language according to the given criteria

Table 1

Rating scale analysis

Score & Rating	Observed count (%)	Observed Average	Andrich Threshold	Outfit MNSQ
1 = memorized	42 (16%)	-4.46	None	0.89
2 = guided	114 (44%)	-1.75	-3.98	0.96
3 = responsively adapted	89 (34%)	0.72	-0.47	0.96
4 = Spontaneously Improvised	13 (5%)	3.63	4.45	1.13

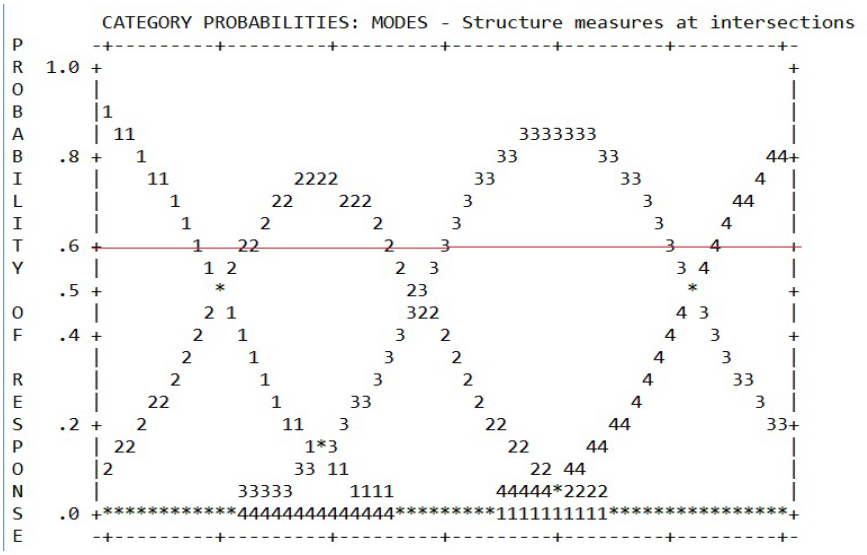

Figure 1. The item response probability of the instrument.

Table 2

Summary statistics of self-assessment speaking rubric

	Mean	Standard Deviation	Strata	Reliability	Cronbach's Alpha Value
Person	-1.07	2.33	3.41	0.84	0.83
Item	0.00	1.27	5.45	0.94	

was slightly low. The person strata (3.41) was categorized as good while the item strata 5.45 was excellent. The bigger value of person and item strata indicated good quality of instrument because it could measure various groups of respondents and items (Sumintono & Widhiarso, 2015). The value of person strata (3.41) enables us to categorise the respondents into three groups of low, average and high ability students.

The person and item reliability as shown in Table 2 was 0.84 and 0.94 respectively, implying that the respondents' responses were consistent and the rubric items were

highly reliable. Cronbach's alpha value of the self-assessment was 0.83 indicating that the overall interaction between respondents and items was highly reliable (Cohen et al., 2007).

Item Fit

The following Table 3 describes the item fit of the six items in the self-assessment speaking skill rubric:

Table 3 shows the Outfit Mean Square (MNSQ), Outfit Z Standard (ZSTD) and also Point Measure Correlation (Pt Meas Corr) of the items in the self-assessment speaking rubric. These three criteria can be used to examine whether or not the items are fit with the model with the value of the Outfit MNSQ that should be in the range 0.5<MNSQ<1.5, Outfit ZSTD value must be between -2.0<ZSTD<+2.0 and the Pt Measure Corr which must be in the range of 0.4<Pt Measure Corr<0.085 (Boone et al., 2014).

The value of items' Outfit MNSQ are all acceptable according to the above mentioned criteria, as well as the value of Outfit ZSTD. There were no items with nearly zero or negative value of Point Measure Correlation indicating there was no item polarity in the rubric or no problematic item which was inconsistent with the construct (Bond & Fox, 2015). Those results showed that all items in the self-assessment speaking rubric were good. It means that all items were valid and could be understood by the respondents.

Principal Component Analysis

The following Table 4 describes the unidimensionality of the rubric:

Unidimensionality, which can be assessed using the principal component analysis, is one among many important aspects in the Rasch model approach that is aimed at examining whether or not the instrument measures only a single underlying attribute (Bond & Fox, 2015) because a good instrument is an instrument which only measures one single variable. The mentioned attribute intended to be measured in this context is students' speaking skill.

Table 3
Item fit

Item	Measure	Standard Error	Outfit MNSQ	Outfit ZSTD	Pt Measure Corr
4	-0.66	0.30	1.14	0.7	0.71
6	-0.48	0.30	1.16	0.8	0.71
1	-0.57	0.30	1.11	0.6	0.64
2	-0.57	0.30	1.00	0.1	0.72
3	2.84	0.34	0.67	-0.8	0.83
5	-0.57	0.30	0.68	-1.6	0.80

Table 4

Principal component analysis of the rubric

```
TABLE 23.0 Rubric.sav                    ZOU527WS.TXT  Sep  4 16:15 2017
INPUT: 43 PERSON  6 ITEM  REPORTED: 43 PERSON  6 ITEM  4 CATS      WINSTEPS 3.73
-------------------------------------------------------------------------------

     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                             -- Empirical --    Modeled
Total raw variance in observations     =         15.8 100.0%         100.0%
  Raw variance explained by measures   =          9.8  62.1%          61.4%
    Raw variance explained by persons  =          5.8  36.9%          36.5%
    Raw Variance explained by items    =          4.0  25.2%          24.9%
  Raw unexplained variance (total)     =          6.0  37.9% 100.0%   38.6%
    Unexplned variance in 1st contrast =          2.0  12.6%  33.3%
    Unexplned variance in 2nd contrast =          1.3   8.2%  21.7%
    Unexplned variance in 3rd contrast =          1.1   7.2%  19.1%
    Unexplned variance in 4th contrast =          0.9   5.5%  14.5%
    Unexplned variance in 5th contrast =          0.7   4.3%  11.3%
```

Note: The recommended font to be used in the original output table from Rasch analysis is Lucida Console 8.

Table 4 shows the value of both raw variance explained by measures and unexplained variance in the 1st contrast. The value of raw variance explained by measures of the self-assessment rubric (62.1%) has exceeded the minimum requirement of 40% while the unexplained variance in the 1st contrast (12.6%) is less than the maximum value of 15%. It can be said that all the items are going to the same direction which is measuring only one variable as it is intended, and that is speaking skill.

The Wright Map

The following Figure 2 described the distribution of both items of rubric and respondents of the study. The left side of the map shows the distribution of the measured ability of students from high ability students (students number 41 and 42) at the top to low ability students at the bottom of the map (student number 35); whereas the right side of the map shows the item distribution in descending order of difficulty.

Wright and Stone (1999) explained that the arrangement of items corresponded to the arrangement of person. Low ability students were below those high ability students; besides that, the easiest items were placed below the most difficult items. According to the students' opinion on their self-assessment about their own skill in speaking, the most difficult aspect of speaking was vocabulary (Aspect3) whilst the easiest aspect was accuracy (Aspect4).

It could be seen that there are many students (on the left side) who have been placed below the lowest item (on the right), implying that many students rated their speaking ability as low.

DISCUSSION

Examining the validity and reliability of the instrument before conducting the

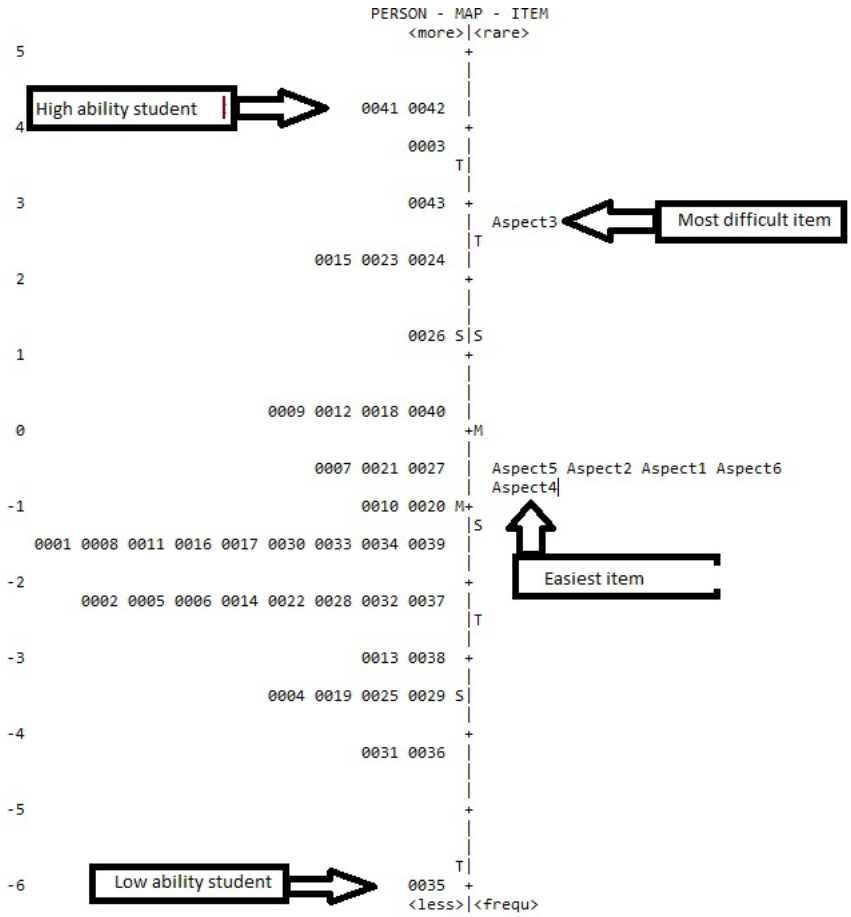

Figure 2. The Wright Map for the instrument

study and collecting the data is very important (Arasinah et al., 2015). The Rasch measurement model approach provided the empirical evidence to prove the quality of the instrument by providing the empirical evidence. The quality of the self-assessment speaking rubric developed by Montgomery in 2000 was examined over several stages. The rubric consists of six items with a four-point rating for each item.

The rating scale analysis showed that respondents were able to clearly differentiate all four rating scales as provided in the rubric by looking at several criteria such as value of observed count and average, Andrich Threshold, Outfit Mean Square and the rating scale peak that exceeded 60% meaning that there is no rating scale needed to be collapsed. To sum up, the four-rating scale category could be continuously used in this rubric.

Person reliability showed high value which indicated the consistency of respondents in giving the response or answer

to the items given. At the same time, the item reliability was also high which indicated good quality of items. The bigger value of person strata showed that the instrument could measure various different groups of respondents. The person strata could also be used to divide the respondents into three different ability levels, namely high ability, average ability and low ability students. It could also be used to classify the items difficulty, such as difficult items and easy items.

Item fit explained whether or not the items carry its function to measure the variable (Sumintono & Widhiarso, 2015) by referring to the Outfit Mean Square, Outfit ZSTD and Point Measure Correlation of items. The result showed that all the values of each item fell in the accepted range of Outfit Mean Square, Outfit ZSTD and Point Measure Correlation, meaning that all items were fitted to the model and those six items successfully carried its function as to measure the speaking skill. It could be concluded there is no misconception with the items and all the items could be understood by the respondents.

Meanwhile, the principal component analysis showed that the value of raw variance explained by measure (62.1%) had exceeded the minimum requirement of 40% and the value of unexplained variance in the 1st contrast did not exceed 15%. This indicated that all the items in the rubric had successfully measured the same variable, as it was intended to measure, which was speaking skill (Aziz et al., 2013).

The last one was the analysis of the Wright map which illustrated the distribution of students' ability in speaking Arabic language based on some certain criteria as stated in the rubric as well as the difficulty level of each aspect of speaking skills. The most difficult aspect of speaking was vocabulary, whereas the easiest one was accuracy. Many students rated themselves as average and low level in speaking, because more students were put at the bottom of the map compared to those who were placed at the top of it. The distribution of items at the right side of the map showed that the items were not evenly spread from the top to bottom. This indicated that the items were slightly too difficult to be mastered according to the students' perception.

The findings of this study might help the teachers to identify how learners assess their own ability. This might give important information to the foreign language teachers as well about what is needed to be improved in the teaching and learning of speaking skill. Giving authority to the learners to assess their ability would give them responsibility for their learning besides preparing them to be the future language teachers who have good judgment skills to assess their students' speaking skills.

CONCLUSION

It could be concluded that the four-point rating scale categories used in this rubric were clearly understood by the respondents by looking at its calibration. The rubric has a good quality to measure speaking

skill by referring to the summary statistics, item fit and principal component analysis. The Cronbach's alpha value as well as the item and person reliability indicated high reliability which means that the rubric was appropriate for use by learners to assess their own speaking ability. Foreign language teachers may also consider to employ this self-assessment speaking rubric in their teaching and learning in order to identify the students' perception about their ability. Furthermore, both teachers and students will be able to discover what should be improved in their teaching and learning process in order to achieve the learning goals.

ACKNOWLEDGEMENT

We would like to thank Mr. Bambang (Lecturer at Universiti Malaya) for helping us with great comments about the Rasch measurement model. and Mr. Burhan Yusuf (IAIN Salatiga, Indonesia) for his assistance during the research process.

REFERENCES

Arasinah, K., Bakar, A. R., Ramlah, H., Soaib, A., & Zaliza, H. (2015). Using Rasch model and confirmatory factor analysis to assess instrument for clothing fashion design competency. *International Journal of Social Science and Humanity, 5*(5), 418-421.

Ary, D., Jacobs, L. C., Razavieh, A., & Sorensen, C. (2006). *Introduction to research in education* (7th ed.). Belmont, USA: Wadsworth.

Aziz, A. A., Masodi, M. S., & Zaharim, A. (2013). *Asas model pengukuram Rasch: Pembentukan skala & struktur pengukuran.* Bangi, Malaysia: Universiti Kebangsaan Malaysia.

Bachman, L. F., & Palmer, A. S. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing, 6*(14), 14-29.

Blanche, P., & Merino, B. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning*, *39*(3), 313-338.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, USA: Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences.* Netherlands: Springer Netherlands.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. New York, USA: Routledge.

Creswell, J. W. (2012). *Educational research: Planning, conducting and evaluating quantitative and qualitative research.* Boston, USA: Pearson.

Dickinson, L. (1987). *Self-intruction in language learning.* New York, USA: Cambridge University Press.

Gardner, D. (2000). Self-assessment for autonomous language learners. *Links & Letters, 7,* 49-60.

Leger, D. d. S. (2009). Self-assessment of speaking skills and participation in a foreign language class. *Foreign Language Annals*, *42*(1), 158-178.

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

Montgomery, C. (2000). *Speaking rubric by Bill Heller*. Retrieved March 20, 2017, from http://languagelinks2006.wikispaces.com/Rubrics

Richards, J. C. (2006). *Communicative language today.* USA: Cambridge University Press.

Srivastava, S. R. (2014). Accuracy vs fluency in English classroom. *New Man International Journal of Multidisciplinary Studies*, *1*(4), 55-58.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan*. Cimahi, Indonesia: Trim Komunikata.

Szyszka, M. (2011). Foreign language anxiety and self-perceived English pronunciation competence. *Studies in Second Language Learning and Teaching* (SSLLT), *1*(2), 283-300.

Wright, B., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Delaware, USA: Wide Range.

Zhang, F., & Yin, P. (2009). A study of pronunciation problems of English learners in China. *Asian Social Science, 5*(6)*,* 141-146.