

ANALYSIS OF MEASURES ITEMS IN DEVELOPMENT OF INSTRUMENTS SELF-ASSESSMENT (RASCH MODELING APPLICATION)

Lailiyah¹

¹Zamzam Syifa Boarding School

Yetti Supriyati², and Komarudin³

²³State University of Jakarta

Lailiyah011@gmail.com

ABSTRACT

This analysis aims to determine the quality of the instrument items that have been developed in the empirical test phase one. Tests were carried out on 46 items to 219 respondents in SMA Ksatria Jakarta. The item quality is seen from the fit or not fit and the level of difficulty of the item that has been developed. The fit or unfit criteria are seen in INFIT and OUTFIT, both MNSQ and ZSTD, and Pt-Measure Correlation values. The level of difficulty of the item is seen in the entry number column which is indicated by the magnitude of the logit value and has been sorted from the hardest to the easiest. Based on the results of analysis with the help of software winstep obtained 39 items statement fit with the model and the number of respondents 194, the three criteria above (MNSQ, ZSTD, and Pt.Measure Correlation) has been met. This means that 39 items are valid. The result of the analysis also shows the most difficult item sequence is item 5 with logit value 63,32, and the easiest item is item 44 with logit value 36,13. The resulting fit instrument must have gone through several stages of analysis. When there are items that are not fit, the item is issued, as well as the respondent. So that obtained a set of measuring instruments that are valid / fit with the model and can be used for the purposes of assessment.

Keywords: *self-assessment, infit, outfit, ZSTD, and Rasch Model.*

INTRODUCTION

The opinions of experts and researchers discussing self-assessment vary. The self-assessment concept in the literature can be summarized into three ways, namely: 1) self-assessment is considered as a personal ability or skill to evaluate a person's or student's knowledge, skills, and performance. 2) self-assessment is used as one type of summative assessment. 3) self-assessment for the purpose of formative assessment acts as a learning strategy or process to improve the quality of student learning (Ziyan: 2016).

Self-assessment is an assessment of learning that if embedded in students how to do self-assessment in the practice of learning, it can be very effective at motivating students to keep moving forward in their own learning (Jayne Bartlett, 2015: 149-150). Consists of two highly related skills, namely self-regulated and self-reflection (Dana s Dunn et al., 2004: 103). Self-assessment is an innovative assessment in improving learning (Benny.A Pribadi, 2009). Involving students to monitor and assess the process and learning outcomes (Burgess A. Angus: 2009).

Self-assessment is currently a major component in the concept of assessment in the classroom, especially in formative assessment. Using self-assessment techniques gives teachers more time to plan their next learning better or more intensively with small groups of students. Because of the results of the self-assessment the teacher can find out more about the development of student

learning, especially the weaknesses and strengths of their students (Siobhan Leahy, 2005: 19-24).

Self-assessment can involve two descriptions, which are about the characteristics of self-employment and evaluation to find out how good the job is and how valuable it is. In this case the accuracy of self-assessment can be determined by comparing students' self-assessments with assessments made by teachers or peers (Gavin T.L. Brown, 2015). The explanation of these two opinions when compared with previous opinions has in common, basically self-assessment is done to improve learning and the quality of student learning outcomes. But both opinions also suggest that the positive impact of self-assessment is felt by the teacher, because the teacher can find out more information about his students. In addition, teachers have time to better prepare for the next teaching and learning process when students do self-assessment. In this case to note is the extent to which the truth of self-assessment, which means to know the honesty of students is not an easy thing, so it is necessary to do the assessment between friends as a comparison.

Therefore, more studies related to self-assessment are needed because the positive impact of self-assessment is quite positive. Self-assessment is one of the non-test assessment techniques conducted by students themselves so that students can monitor themselves, and know what aspects have been and have not been known related to the learning. Students identify the extent to which their learning outcomes are achieved, and determine whether they are good or must be corrected. Students target the achievement of subsequent learning and determine how to achieve it, and students identify weaknesses and their excesses with self-reflection.

Associated with the process of preparing instrument items the author uses a modern approach namely modeling rasch. Rasch modeling for the first time by Dr. Georg Rasch is a mathematician from Denmark. In the 1950s. Georg is faced with an analysis of the results of the examination of elementary students at different grades. The exam questions are used the same and are not based on the material according to the class. This is where the discovery of model rasch begins that begins with the idea of the problem facing it (Bambang Sumintono, 2015: 35). Rasch measurement is quantitative but also qualitative. Researchers who use rasch measurement are aware that the measurement results that have been carried out require qualitative reflective, not just a number of numbers that appear after the analysis using Winstep software. Analysis using modeling gets quite a lot of information such as person map items, statistical items: misfit order, item statistics: measure order, scalogram, person statistics: misfit person, person statistics: measure order, unidimensionality, statistical summary, and still more. (William J. Boone et al., 2014). But in this article only the item measure analysis is discussed. Item measure is one of the analysis results that inform the fit or not a statement and grain difficulty points indicated premises logitnya magnitude.

METHODOLOGY

Methodology in this research uses a modern approach to modeling rasch (item respon theory). The research design was descriptive with the research subjects 219 respondents' response patterns to the instruments given, which were 46 items. The study was conducted at the Ksatria high school in Jakarta. selection of sample locations based on school accreditation, namely schools with an A

accreditation. After the data is collected, it is then analyzed quantitatively with winstep software. The analysis was carried out several times until the results of the analysis were fit with the measurement model, namely rasch modeling. After that, the results are interpreted qualitatively and support with tables and graphs to facilitate the reader.

RESULTS

Analysis of the quality of the items in the instrument can be seen in the statistics item table: item measure to find out the items are valid or not. To know the validity of the instrument by looking at three criteria ie infit value and outfit mean square, infit and outfit ZSTD, and Pt.Measure Correlation with the following criteria.

Interpretation of parameter-level mean-square fit statistics	
> 2.0	<i>Distorts or degrades the measurement system</i>
1.5 – 2.0	<i>Unproductive for construction of measurement, but not degrading</i>
0.5 – 1.5	<i>Productive for measurement</i>
< 0,5	<i>Less productive for measurement, but but not degrading. May produce misleadingly good reliabilitas and separation</i>
Interpretation of mean-square fit statistic values (reprinted with permission from wright &linacre, 1994)	
Standardized value	Implication for measurement
≥ 3	<i>Data very unexpected if they fit the model (perfectly), so they probably do not. But, with large sample size, substantive misfit may be small.</i>
2.0 – 2.9	<i>Data noticeably unpredictable</i>
-1.9 – 1.9	<i>Data have reasonable predictability</i>
≤ -2	<i>Data are too predictable. Other “dimension” may be constraining the response patterns.</i>
Guidelines for the interpretation of ZSTD values from Linacre (2002)	

Based on the above criteria table will be known items that are valid and invalid (fit with the model). The following table item statistic: measure order, from the analysis of 46 items and 219 respondents. This analysis provides the fit or non-fit information of an items, as well as the degree of difficulty of the item.

TABLE 13.1 C:\Users\Laili\Documents\FINAL UJI EM ZOU643WS.TXTE Jul 9 5:44 2018

INPUT: 219 Person 46 Item REPORTED: 219 Person 46 Item 5 CATS
WINSTEPS 3.73

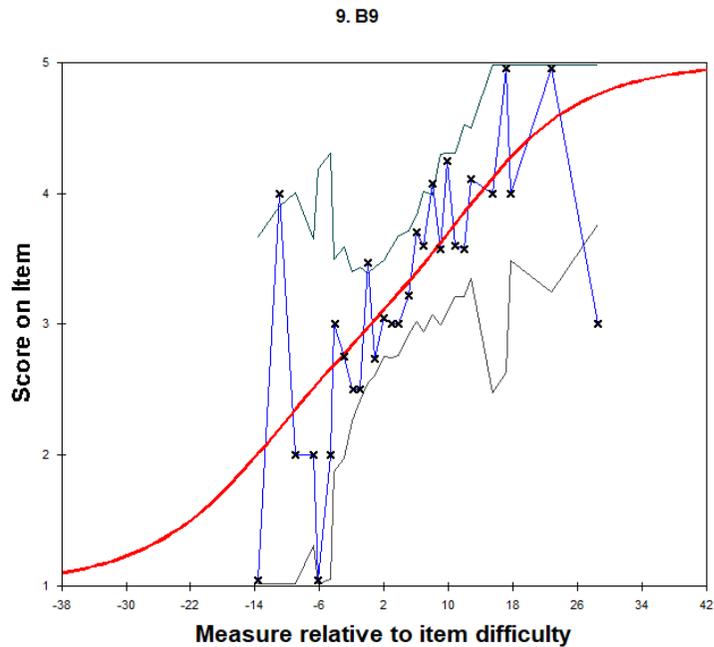
Person: REAL SEP.: 2.71 REL.: .88 ... Item: REAL SEP.: 7.29 REL.: .98

Item STATISTICS: MEASURE ORDER

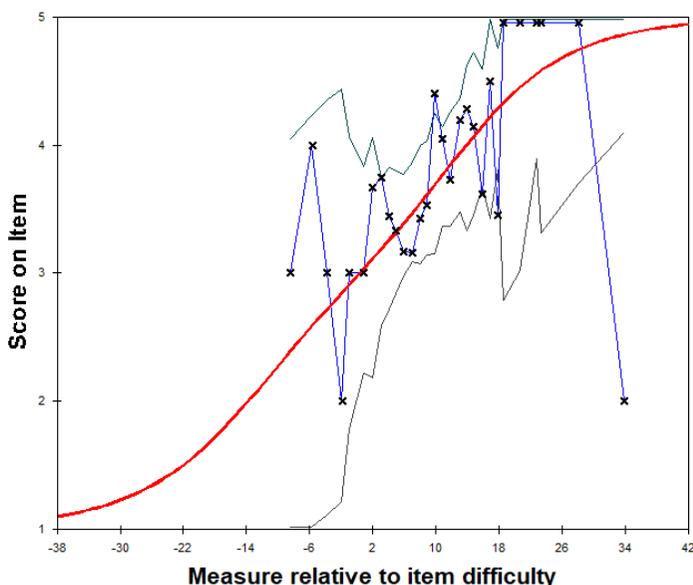
ENTRY	TOTAL	TOTAL	MODEL	INFIT	OUTFIT	PT-MEASURE								
			EXACT MATCH											
	NUMBER	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	Item
5	553	219	59.88	.81	.84	-1.8	.84	-1.8	.52	.43	48.9	44.3	B5	
12	553	219	59.88	.81	.84	-1.8	.84	-1.8	.52	.43	48.9	44.3	B12	
21	553	219	59.88	.81	.84	-1.8	.84	-1.8	.52	.43	48.9	44.3	B21	
23	553	219	59.88	.81	.84	-1.8	.84	-1.8	.52	.43	48.9	44.3	B23	
27	553	219	59.88	.81	.84	-1.8	.84	-1.8	.52	.43	48.9	44.3	B27	
42	553	219	59.88	.81	.84	-1.8	.84	-1.8	.52	.43	48.9	44.3	B42	
20	595	219	57.16	.81	1.44	4.0	1.44	4.0	.29	.43	39.3	47.9	B20	
34	606	218	56.23	.81	.99	-.1	.98	-.1	.35	.43	52.8	49.1	B34	
29	612	219	56.06	.81	.89	-1.2	.89	-1.1	.43	.43	57.1	49.1	B29	
1	628	219	55.01	.81	.91	-.9	.90	-1.0	.41	.43	55.3	51.1	B1	
7	628	219	55.01	.81	.91	-.9	.90	-1.0	.41	.43	55.3	51.1	B7	
8	628	219	55.01	.81	.91	-.9	.90	-1.0	.41	.43	55.3	51.1	B8	
14	628	219	55.01	.81	.91	-.9	.90	-1.0	.41	.43	55.3	51.1	B14	
2	638	219	54.36	.81	1.15	1.5	1.15	1.5	.31	.43	53.0	51.7	B2	
3	638	219	54.36	.81	1.15	1.5	1.15	1.5	.31	.43	53.0	51.7	B3	
6	638	219	54.36	.81	1.15	1.5	1.15	1.5	.31	.43	53.0	51.7	B6	
19	638	219	54.36	.81	1.15	1.5	1.15	1.5	.31	.43	53.0	51.7	B19	
35	638	219	54.36	.81	1.15	1.5	1.15	1.5	.31	.43	53.0	51.7	B35	
33	664	219	52.67	.81	1.19	1.8	1.18	1.7	.30	.43	45.2	51.4	B33	
13	681	219	51.57	.80	1.41	3.7	1.41	3.7	.42	.43	37.4	50.7	B13	
9	724	219	48.81	.80	1.30	2.9	1.34	3.2	.42	.43	35.2	48.2	B9	
16	724	219	48.81	.80	1.30	2.9	1.34	3.2	.42	.43	35.2	48.2	B16	
11	726	219	48.68	.80	.99	.0	1.02	.2	.27	.43	49.8	48.1	B11	
41	730	219	48.43	.80	1.11	1.2	1.16	1.6	.32	.43	50.7	47.8	B41	
45	749	219	47.23	.79	1.16	1.6	1.16	1.6	.34	.43	40.2	45.4	B45	
4	767	219	46.10	.79	.92	-.9	.91	-1.0	.60	.43	45.7	43.4	B4	
25	767	219	46.10	.79	.92	-.9	.91	-1.0	.60	.43	45.7	43.4	B25	
28	767	219	46.10	.79	.92	-.9	.91	-1.0	.60	.43	45.7	43.4	B28	
30	767	219	46.10	.79	.92	-.9	.91	-1.0	.60	.43	45.7	43.4	B30	
31	767	219	46.10	.79	.92	-.9	.91	-1.0	.60	.43	45.7	43.4	B31	
32	767	219	46.10	.79	.92	-.9	.91	-1.0	.60	.43	45.7	43.4	B32	
36	767	219	46.10	.79	.92	-.9	.91	-1.0	.60	.43	45.7	43.4	B36	
37	767	219	46.10	.79	.92	-.9	.91	-1.0	.60	.43	45.7	43.4	B37	
39	767	219	46.10	.79	.92	-.9	.91	-1.0	.60	.43	45.7	43.4	B39	
43	767	219	46.10	.79	.92	-.9	.91	-1.0	.60	.43	45.7	43.4	B43	
10	774	219	45.66	.79	.73	-3.3	.74	-3.1	.35	.43	50.7	42.9	B10	
18	774	219	45.66	.79	.73	-3.3	.74	-3.1	.35	.43	50.7	42.9	B18	
15	796	219	44.29	.79	.91	-1.0	.91	-1.1	.35	.42	49.8	40.9	B15	
22	796	219	44.29	.79	.91	-1.0	.91	-1.1	.35	.42	49.8	40.9	B22	
24	796	219	44.29	.79	.91	-1.0	.91	-1.1	.35	.42	49.8	40.9	B24	
38	796	219	44.29	.79	.91	-1.0	.91	-1.1	.35	.42	49.8	40.9	B38	

26	808	219	43.54	.79	1.27	3.0	1.47	4.9	.27	.42	40.6	40.1	B26
46	823	219	42.60	.79	1.09	1.1	1.10	1.2	.33	.42	32.4	38.3	B46
17	876	219	39.20	.81	.99	.0	1.04	.5	.32	.40	42.5	35.9	B17
40	876	219	39.20	.81	.99	.0	1.04	.5	.32	.40	42.5	35.9	B40
44	876	219	39.20	.81	.99	.0	1.04	.5	.32	.40	42.5	35.9	B44
-----+-----+-----+-----+-----+-----													
MEAN	706.3	219.0	50.00	.80	1.00	-.1	1.00	-.1			47.4	45.3	
S.D.	94.9	.1	6.08	.01	.17	1.7	.18	1.8			5.7	4.5	

Based on the above analysis, it is known that 7 items do not meet the fit criteria. They are number of 20 (4,0), 9 (2,9), 16 (2,9), 13 (3,7), 10 (-3,1), 18 (-3,1), and 26 (4,9). The seven items have ZSTD value greater than 2 and less than -2 which means data can not be predicted. Item nonconformities can also be seen in ICC's expected score pattern graphs, such as points 9 and 26 as follows.



26. B26



The graph above informs that there is a response pattern that is too far with the ideal model line curve. Therefore, further analysis must be carried out, namely by issuing items that are not fit in a row until the results of the analysis are obtained where all items fit the model. In the advanced analysis also issued respondents who are not fit as many as 25 respondents. So that the results of the analysis are fit with the following model.

TABLE 13.1 C:\Users\Laili\Documents\FINAL UJI EM ZOU117WS.TXTE Mar 9 6:55
2018
INPUT: 194 Person 39 Item REPORTED: 194 Person 39 Item 5 CATS WINSTEPS
3.73

Person: REAL SEP.: 2.73 REL.: .88 ... Item: REAL SEP.: 8.39 REL.: .99

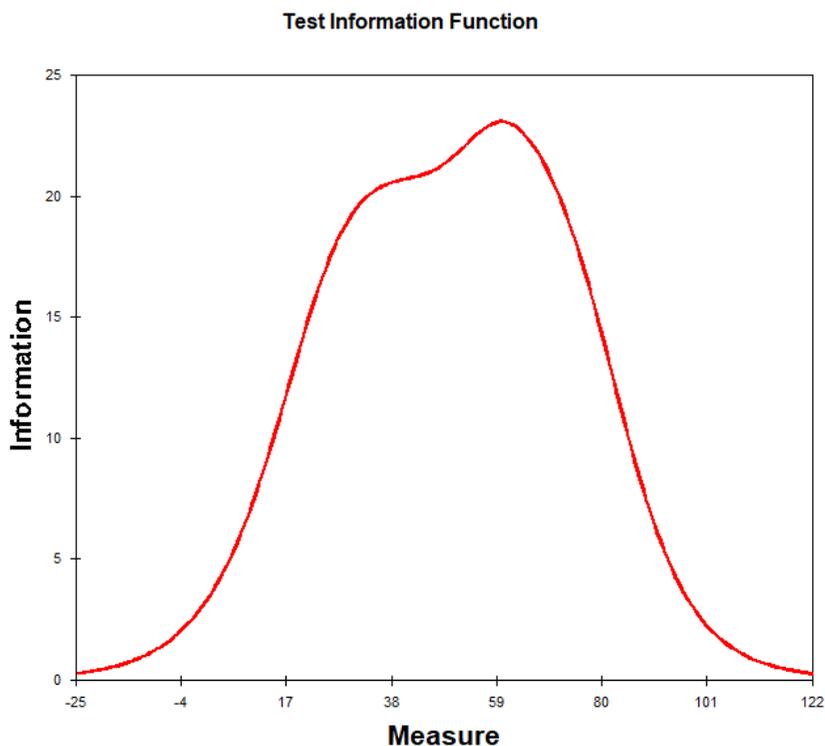
Item STATISTICS: MEASURE ORDER

ENTRY	TOTAL	TOTAL	MODEL	INFIT	OUTFIT	PT-MEASURE
[EXACT MATCH]						
[NUMBER	SCORE	COUNT	MEASURE	S.E.	[MNSQ ZSTD]	[MNSQ ZSTD]
[CORR. EXP.] OBS% EXP% Item						
5	482	194	63.32	.97 .85	-1.5 .85	-1.6 .51
10	482	194	63.32	.97 .85	-1.5 .85	-1.6 .51

15	482	194	63.32	.97	.85	-1.5	.85	-1.6	.51	.45	54.6	49.8	B21
17	482	194	63.32	.97	.85	-1.5	.85	-1.6	.51	.45	54.6	49.8	B23
20	482	194	63.32	.97	.85	-1.5	.85	-1.6	.51	.45	54.6	49.8	B27
35	482	194	63.32	.97	.85	-1.5	.85	-1.6	.51	.45	54.6	49.8	B42
22	537	194	57.99	.99	1.13	1.2	1.15	1.3	.42	.45	58.8	55.5	B29
27	539	193	57.47	1.00	1.18	1.6	1.17	1.5	.37	.45	55.4	55.9	B34
1	557	194	56.01	1.00	1.09	.8	1.09	.8	.39	.45	53.6	57.5	B1
7	557	194	56.01	1.00	1.09	.8	1.09	.8	.39	.45	53.6	57.5	B7
8	557	194	56.01	1.00	1.09	.8	1.09	.8	.39	.45	53.6	57.5	B8
11	557	194	56.01	1.00	1.09	.8	1.09	.8	.39	.45	53.6	57.5	B14
2	563	194	55.42	1.00	1.28	2.0	1.30	2.0	.32	.45	54.1	57.9	B2
3	563	194	55.42	1.00	1.28	2.0	1.30	2.0	.32	.45	54.1	57.9	B3
6	563	194	55.42	1.00	1.28	2.0	1.30	2.0	.32	.45	54.1	57.9	B6
14	563	194	55.42	1.00	1.28	2.0	1.30	2.0	.32	.45	54.1	57.9	B19
28	563	194	55.42	1.00	1.28	2.0	1.30	2.0	.32	.45	54.1	57.9	B35
26	572	194	54.53	1.00	1.53	1.0	1.52	1.9	.28	.45	44.8	57.5	B33
9	642	194	47.83	.96	1.14	1.3	1.12	1.1	.30	.46	50.5	53.7	B11
38	646	194	47.47	.95	1.36	1.2	1.37	1.1	.35	.46	42.3	52.8	B45
34	655	194	46.66	.95	1.14	1.3	1.17	1.5	.39	.46	46.9	51.7	B41
4	686	194	43.94	.93	.75	-1.9	.73	-2.0	.68	.47	51.5	46.9	B4
19	686	194	43.94	.93	.75	-1.9	.73	-2.0	.68	.47	51.5	46.9	B25
21	686	194	43.94	.93	.75	-1.9	.73	-2.0	.68	.47	51.5	46.9	B28
23	686	194	43.94	.93	.75	-1.9	.73	-2.0	.68	.47	51.5	46.9	B30
24	686	194	43.94	.93	.75	-1.9	.73	-2.0	.68	.47	51.5	46.9	B31
25	686	194	43.94	.93	.75	-1.9	.73	-1.0	.68	.47	51.5	46.9	B32
29	686	194	43.94	.93	.75	-1.9	.73	-1.0	.68	.47	51.5	46.9	B36
30	686	194	43.94	.93	.75	-1.9	.73	-1.0	.68	.47	51.5	46.9	B37
32	686	194	43.94	.93	.75	-1.9	.73	-1.0	.68	.47	51.5	46.9	B39
36	686	194	43.94	.93	.75	-1.9	.73	-1.0	.68	.47	51.5	46.9	B43
12	708	194	42.07	.92	.97	-.3	.98	-.1	.39	.47	49.5	44.0	B15
16	708	194	42.07	.92	.97	-.3	.98	-.1	.39	.47	49.5	44.0	B22
18	708	194	42.07	.92	.97	-.3	.98	-.1	.39	.47	49.5	44.0	B24
31	708	194	42.07	.92	.97	-.3	.98	-.1	.39	.47	49.5	44.0	B38
39	722	194	40.90	.91	1.22	1.4	1.25	1.5	.32	.47	33.0	42.3	B46
13	779	194	36.13	.92	1.03	.4	1.09	1.0	.30	.46	43.8	41.1	B17
33	779	194	36.13	.92	1.03	.4	1.09	1.0	.30	.46	43.8	41.1	B40
37	779	194	36.13	.92	1.03	.4	1.09	1.0	.30	.46	43.8	41.1	B44
-----+-----+-----+-----+-----+-----													
MEAN	622.5	194.0	50.00	.96	1.00	-.2	1.00	-.2			51.0	50.2	
S.D.	91.3	.2	8.47	.03	.21	2.1	.22	2.1			4.7	5.6	

Based on the above table it is known that the difficulty level of the items shown in the measure column and sorted from the hardest item to the easiest point. The most difficult item is item 5 with a 63.32 logit value. Then point 12, 21, and so on until the easiest ie item 44 with a logit value of 36.13. Another information is the following ICC graph which shows that the item statement can produce optimal

information if given to respondents who have medium ability.



CONCLUSION

Analysis of item measure on testing of empirical instrument of stage one is executed with total item 46 and respondent 219. Based on result of analysis obtained that item instrument developed according to model after done several times analysis, that is by issuing outlier respondents counted 25 people and item invalid as many as 7 items. So that the total number of fit 39 points. Instrument can provide more accurate measurement results if given to medium-ability respondents.

ACKNOWLEDGEMENT

This research was supported by the Education Fund Management Institute (LPDP) of the Ministry of Finance of the Republic of Indonesia. Thanks to state university of jakarta and the promoters who have given support so that this article can be resolved properly.

REFERENCES

- Bartlett, Jayne *Outstanding Assessment for Learning in The Classroom*. Routledge. New York. 2015. Hal 149-150.
- Benny, A. Pribadi. *Model Desain Sistem Pembelajaran*. Jakarta. Dian Rakyat. 2009.

- Boone, William J & John R. Staver, Dkk. *Rasch Analysis in the human Sciences*. 2013.
- Angus. Burgess A. Jean dan Green. *Digital Media dan society series*. Cambridge.Policy Press. 2009.
- Brown,Gavin T.L. Heidi L. Andrade & Fei Chen. *Accuracy in Student Self-Assessment: Directions and Cautions for Research*. Assessment in Education: Principles, Policy &Practices, 2015.
- Dunn,Dana s Chandra M. Mehrotra, dan Jane S. Halonen. *Measuring Up Educational Assessment Challenges and Practices for Psychology*. American Psychological Association. Washington Dc. 2004. Hal.173.
- Leahy, Siobhan et. al., *Classroom Assessment: Minute By Minute, Day By Day*. Educational Leadership. Vol 63. No. 3. Novemver 2005. P.19-24.
- Sumintono Bambang, Wahyu Widiyatmoko. *Aplikasi Pemodelan Rasch Pada Assessment Pendidikan*.Trim Komunikata:2013.
- Yan.Zi*The Self-Assessment Practices Of Hongkong Secondary Student: Findings With A New Instrument*. Jurnal Of Applied Measurement. 21 November 2016.