The comparison of descriptive statistical parameter estimation stability using raw scores and rasch model

The comparison of descriptive statistical parameter estimation stability using raw scores and rasch model

P Susongko *

Universitas Pancasakti Tegal

*Coresponding author: purwosusongko@upstegal.ac.id

Abstract. This research aims at analyzing the comparison of descriptive statistical Parameter Estimation stability using raw scores and Rasch model. The empirical data were the responses of the 12th Grade Science Students of Senior High School on Science Literary Test based on the integrated mathematics and natural sciences conducted at *SMAN 2* and *SMAN 3* Tegal, Central Java. This research employed a bootstrapping method assisted with SPSS version 21, while the Rasch model was assisted with R program version 3.6.3 eRm package Version 1.0.1. The parameter stability estimation was seen from its error standard and bias scores. The scores using the Rasch model was proven giving higher stability when compared to that using the raw scores in its descriptive statistical parameter estimation both from its error standard and bias aspects. Based on the error standard used, it showed that the mean and the standard deviation estimation when using the Rasch model scores was around 8 times more stable when compared to that when using the raw scores, while its median estimation were 16-18 times. Due to the use of bias measurement, it showed that the mean estimation when using the Rasch model scores was 6-10 times more stable when compared to that using the raw scores, while its median estimation and standard deviation were respectively 43-282 times and 7-10 times more stable.

1. Introduction

The use of raw scores to measure achievements basically has some weaknesses: (a) raw scores are basically not the measurement results. Raw scores are more precisely the number of correct answers to the questions; (b) raw scores are the preliminary information. Raw scores are usually stated in percentage (%) as the summary of numerical data; (c) raw scores have a weak quantitative significance. The meaning of quantitative from the obtained raw scores will be different depending on the number of questions, while the percentage of the correct answers always depends on the questions' difficulty levels; (d) raw scores do not show someone's ability on certain tasks. Raw scores cannot explain more about the problem difficulties; and (e) raw scores and the percentage of correct answers are not always linear. In the linear test, students with score 15 (scale 0 to 100) always has a higher ability than those with score 10. However, both empirically have the same ability [1,2].

Van Zile-Tamsen state that the raw scores approach or classical test theory (CCT) has some limitations: (1) in fact, the obtained scores depend on the samples and bias to the actual scores; (2) inability to resolve the missing data, (3) Measuring the reliability seen only from the size of Cronbach's alpha coefficient, while the validity is proven only based on the test content and the correlation of scores from a scale with the other measurement results; and (4) It is very difficult to determine the effectiveness of items on the target population and its contributions to the latent construct overall measurement [3].

According to Mok and Wright, an objective measurement concept in social and educational assessment sciences should have 5 criteria: (1) resulting in a linier size with the same interval; (2) having a precise estimation process, (3) Identifying the misfit and outlier items; (4) Having the ability to resolve the missing data; and (5) resulting in independent measurements from the parameters under study [4]. From those 5 criteria, only Rasch model meets, so far, the requirements. The measurement quality in the educational assessment conducted using the Rasch model will have the same quality as that conducted in physical dimensions in the field of Physics [5]. In measuring the modern test theory, Rasch model is considered as the most objective one. The use of Rasch model in educational assessment has the excellent in specific objectivity and high item parameter stability estimation [6].

Rasch model connects the correctly answering opportunities in each item (P(θ)) as the function of ability (θ) with the item's constant difficulty level (b) through the relationship as shown in equation 1.

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad (1)$$

The Rasch model has been further separately developed from IRT and even developed wider in the polytomous scoring. The implementation of Rasch model in learning achievement since introduced by the inventor, Georg Rasch, in 1960 is recently used extensively used not only in educational but also in medical and public health world [7,8,9].

In educational and social research, the scores resulted are still in the ordinal or discrete measurement level since the scores obtained from questionnaires or tests in educational or social researches are calculated by summing the number of correct answers in the form of polytomous and dichotomous. Up to now, those scores are considered as intervals which are then analyzed using the parametric statistics. Some experts give some formulas to convert or transform the ordinal to interval scale. Edward, & Thurstone have developed the simplest technique by using MSI (Method of Successive Interval) [10]. However, the transformation principally creates the essential questions related to the transformation principles in which the lower scale is impossible to be transformed to the higher one. The higher score can be transformed to the simpler scale, yet not in contrary.

The Rasch scoring answers the problems. The scoring using Rasch model results in scores in the interval level [11,12,13,14]. By conversing the raw scores to Rasch scores, it is expected that those scores have meet the interval assumption as the requirement for the parametric analysis [15,16]. In its implementation to the statistical parameter estimation, both scorings require the stability level test. Thus, a research on how far the statistical parameter estimation stability in both raw scores and Rasch scores is necessary to conduct. This research aims at examining the comparison between the descriptive statistical parameter estimation stability in raw scores and that in Rasch scores.

2. Method

This research used the empirical data in the form of responses from the XII Grade Natural Science Students of Senior High School on the Science Literary Test which was based on the integrated mathematics and natural sciences conducted at *SMAN 2* Tegal and *SMAN 3* Tegal. The implementation at SMA N 2 was on 11 February 2020 at 7.30 a.m. to 9.30 a.m. of Western Indonesian Zone, while the tests conducted at SMA N 3 was on 21 February 2020 at 8 a.m. to 10 a.m. of Western Indonesian Zone. 310 participants were involved in this research with the distributions shown in Table 1. Those 310 samples were selected as the inference bases as the bootstrapp analysis would not give good results when using too small samples [17].

Table 1. Distributions of Participants Taking the Science Literary Test based on The Integrated Mathematics and Natural Sciences

No	School	Male Participants	Female Participants	Total
1	SMA N 2 Tegal	33	86	119
2	SMA N 3 Tegal	69	122	191
Total		102	208	310

Bootstrapping was the method used in this research. Bootstrapp method resulted in the statistical inference. The resulted statistical inference was in the form of error standard and bias estimation, confidence interval, and hypothetical test without assumption, such as normal or same variance distributions. Thus, bootstrapp method could be more accurate than the classical assumption based on the normal or t distribution [17, 18].

These research stages started by estimating the descriptive parameter obtained from both raw scores and Rasch model scores of science literary test based on the integrated mathematics and natural sciences. The estimated descriptive parameter included mean, median, and standard deviation. This analysis was assisted by the R program version 3.6.3 eRm package Version 1.0.1. Bootstrapping was then performed to the estimated descriptive parameter from the raw scores and Rasch scores using the new samples which had the same size with the actual samples from 300,600,900,1200,1500,1800 to 2100. The estimation stability could be seen from the bias size and error standard resulted from the parameter estimation based on the bootstrapping analysis assisted by SPSS program Version 21. The bias showed the differences between the mean statistical scores of all bootstrapp samples and actual samples. The error Standard showed the mean error scores within the bootstrapp processes. Lower error standard and bias scores on both (raw and Rasch model) scoring populations showed the higher stability.

3. Result and Discussion

The research result began by displaying descriptive statistical parameter estimation data of both raw scores and Rasch scores from the empirical data of 310 participants as explained in Table 2. Based on the empirical data, resampling was performed from 300, 600, 900, 1200, 1500, 1800 to 2100 times with the same size and the actual samples that resulted in the error standard as shown in Table 3. Similarly, the bias was also shown in Table 4. In Table 2, there were 310 valid students' responses and resulted in three descriptive statistical parameters consisting of mean, median and standard deviation.

Table 2 Descriptive Statistical Parameter of Empirical Data in Both Raw Scores and Rasch Model Scores

Parameter	Raw scores	Rasch model scores
Sample Size	310	310
Mean	21.77	0.14
Median	22.00	0.14
Std. Deviation	4.83	0.57

Table 3 showed that the use of Rasch model scores had lower error standard than that of raw scores in all mean, median, and standard deviation estimations. The highest accuracy in the use of Rasch model scores was in the median estimation in which the mean ratio of error standard in the use of raw scores to the Rasch model scores was 16.39 and concluded that the use of Rasch model scores was 16 to 18 times more stable than the use of raw scores in median estimation seen from the error standard aspect. In estimating the mean and standard deviation, the ratio of the raw standard error score to the Rasch model score has value of about 8. It showed that the use of Rasch model scores provided more accurate analysis result than that of raw scores.

Table 3. Parameter Estimation Error Standard using Bootstrapping in Descriptive Statistics with Raw Scores and Rasch Model Scores

Parameter	Scoring	300	600	900	1200	1500	1800	2100
Mean	Raw Scores	0.2794	0.2709	0.2766	0.2753	0.2806	0.2761	0.2822
	Rasch Scores	0.0331	0.0326	0.0329	0.0328	0.0335	0.0324	0.0334
	Ratio	8.44	8.33	8.41	8.39	8.37	8.52	8.45

Parameter	Scoring	300	600	900	1200	1500	1800	2100
Median	Raw Scores	0.4712	0.4302	0.4599	0.4602	0.4566	0.4371	0.4497
	Rasch Scores	0.0262	0.0299	0.0265	0.0288	0.0286	0.0264	0.0277
	Ratio	18.0	14.4	17.4	16.1	16.0	16.6	16.2
Standard Deviation	Raw Scores	0.17299	0.17299	0.17850	0.17347	0.17128	0.17835	.17750
	Rasch Scores	0.02410	0.02419	0.02497	0.02379	0.02348	0.02479	.02459
	Ratio	7.2	7.2	7.1	7.3	7.3	7.2	7.2

Table 4 Parameter Estimation Bias using Bootstrapping in Descriptive Statistics with Raw Scores and Rasch Model Scores

Parameter	Scoring	300	600	900	1200	1500	1800	2100
Mean	Raw Scores	-0.0208	-0.0135	-0.0102	0.0068	-0.0075	0.0029	0.0131
	Rasch Scores	-0.0023	-0.0021	-0.0015	0.0008	-0.0009	0.0001	0.0012
	Ratio	9	6.5	6.8	8.5	8.3	29	10.9
Median	Raw Scores	-0.2150	-0.1900	-0.2233	-0.1850	-0.1937	-0.1933	-0.1693
	Rasch Scores	-0.0002	-0.0044	-0.0017	-0.0007	-0.0013	-0.0018	-0.0006
	Ratio	107.5	43.2	131.4	264.3	149	107.4	282.2
Standard Deviation	Raw Scores	-0.01889	-0.01845	-0.0243	-0.0091	-0.0221	-0.0168	-0.0072
	Rasch Scores	-0.00273	-0.00184	-0.0031	-0.0012	-0.0032	-0.0019	-0.0007
	Ratio	6.9	10	7.9	7.5	7	8.7	10

Table 4 showed that the use of Rasch model scores had smaller bias than the use of raw scores in the estimation of mean, median and standard deviation. In the mean parameter estimation, the use of Rasch model resulted in bias of between 0.002 to 0.0001, while the use of raw scores resulted in bias of between 0.01 to 0.002. The ratio of bias aspect for the mean parameter estimation, the use of Rasch model had the stability of 6 to 10 times than the use of raw scores. The research results also showed that the use of Rasch model had the tendency that the more the resampling is made, the lower the bias will be.

The stability differences were quite significant in the median parameter estimations. In the median parameter estimations, the use of Rasch model resulted in bias of between 0.004 to 0.0002, while the use of raw scores resulted in bias of around 0.2 that if seen from the ratio, the use of Rasch model had the stability of between 43 to 282 times than the use of raw scores. The result of this research also showed that the resampling of 300 precisely gave the smallest bias in the use of Rasch model, while in the use of raw scores showed that all resampling relatively had the same bias value.

In the standard deviation estimation, using both raw scores and Rasch model scores had the smallest bias value in the resampling process of 2100. The bias in the use of raw scores was between 0.02 to 0.007, while the use of Rasch model scores was between 0.003 to 0.0007. The result of this research was quite interesting in the parameter estimation of standard deviation that the bias ratio in the use of raw scores to the Rasch model scores was relatively constant at 6-10. It showed that Rasch model was 6-10 times more stable than the use of raw scores.

The results of this research were in line with some previous studies related to the comparison of the Rasch model scores with the raw scores. Zhao compared the raw scoring and Rasch scoring in the mastery changing estimation of students' learning achievements [19]. The results of research conducted

by Zhao showed that Rasch model had greater ability to show the students' learning achievements than the raw scores. Thus, Rasch model scores could be generalized to measure the students' learning achievement changes when continuing their study to higher education levels. Van Zile-Tamsen compared the raw scores with the Rasch scores in their implementation to the psychometric aspects [3]. The results research conducted by Van-Zile Tamsen showed that the Rasch model integrated with the professional assessment and measurement context was more effective than using the raw scores in the rating scale development.

The results of this research also answered the weaknesses of raw or classical scores during this time assumed as the ordinal scores. The British Medical Journal has already warned since 1986 that the use of descriptive statistic parameter using the raw scores or those obtained by counting the correct answers was invalid since having no constant interval [11]. The conversion of raw scores in Rasch model scores was seen able to change the ordinal scales in the educational and social assessments to become the interval scales [20,21].

4. Conclusion

The Rasch model scoring was proven giving higher stability than using the raw scores in the descriptive statistical parameter estimation from both error standard and bias aspect. By using the error standard, it showed that the Rasch model scoring made the mean and the standard deviation estimation have around 8 times more stable than using the raw scores. Meanwhile, the estimation of median using Rasch model were 16 to 18 times more stable than using the raw scores. By using the bias measurement, it showed that in the Rasch model scoring, the mean estimation was 6-10 times more stable than using the raw scores, while the median estimation was 43-282 times more stable than using the raw scores. Furthermore, by using the bias measurement, the standard deviation estimation using the Rasch model scoring was 7 to 10 times more stable than using the raw scores.

References

[1] Andrich D & Marais I 2019 *A Course in Rasch Measurement Theory: Measuring in The Educational, Social and Health Sciences* (Singapore: Springer)
[2] Sumintono B 2018 *1st Int. Conf. on Education Innovation vol 1* (Kuala Lumpur: Atlantis Press) 38
[3] Van Zile-Tamsen C 2017 *Res. High. Edu.* **58** (8) 922
[4] Mok M and Wright, B 2004 *Introduction to Rasch Measurement: Theory, Models and Applications* (Minnesota: Jam Press)
[5] Sumintono, B, & Widhiarso, W 2014 *Rasch Model Application For Social Sciences Research* (Cimahi: Trim Komunikata Publishing House)
[6] Wu, M & Adams, R 2007 *Applying the Rasch model to psycho-social measurement: A practical approach* (Melbourne: Educational Measurement Solutions)
[7] Yen-Mou L, Yuh-Yin W, Ching-Lin H, Chih-Lung L, Shiuh-Lin H, Kuang-I C & Yi-Jing, L 2013 *J. Health. Quality . Life. Out* **11 (**119) 1
[8] Smith A B, Fallowfield L J, Stark D P, Velikova G & Jenkins V 2010 J. *Health. Qual. Life. Out.* **8** (45) 1
[9] Ayele D G, Zewotir T, & Mwambi H 2014 *Int.J. Envir. Res. and Pub. Heal.* **11** (7) 6681
[10] Edwards A L & Thurstone L L 1952 *Psycho.* **17** (2) 169
[11] Grimby G, Tennant A. & Tesio L 2012 *J. Rehabil. Med* **44** 97
[12] Dencker A, Sunnerhagen K S, Taft C & Lundgren-Nilsson Å 2015 *Heal. Qual of Life Out.* **13 (**1) 1
[13] Walton D M, Wideman T H & Sullivan M J 2013. *The Clinic. J. of Pain* **29** (6) 499
[14] Medvedev O N, Siegert R J , Mohamed A D , Shepherd D, Landhuis E & Krägeloh, CU 2017. *J. Happ.Stud.* **18** (5) 1425
[15] Norman G 2010 *Advan. Heal. Sci.Edu.* **15** (5) 625
[16] Harpe SE 2015 *Curr. Phar. Teach. Learn*. **7** (6) 836

[17] Hesterberg T 2011 *Bootstrap Wiley. Int. Rev: Com. Stat.* **3** (6) 497
[18] Lemoine F, Entfellner JBD, Wilkinson E, Correia D, Felipe MD, De Oliveira, T & Gascuel O 2018 *Nature* **556** (7702) 452
[19] Zhao Y , Huen JM & Chan YW 2017 *Res. High. Edu*. **58** (6) 605
[20] Medvedev O N, Siegert R J, Feng X J., Billington D R, Jang J Y, & Krägeloh C U 2016 *Mindfulness* **7** (2) 384
[21] Bond T, Yan Z, & Heene M 2020 *Applying the Rasch Model: Fundamental Measurement in The Human Sciences* (New York: Routledge)