International Journal of Applied Mathematics
& Statistical Sciences (IJAMSS)
ISSN(P): 2319–3972; ISSN(E): 2319–3980
Vol. 9, Issue 2, Feb–Mar 2020; 53–62

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

STATISTICAL CONVERSION OF RAW SCORE TO SCALED SCORE THROUGH BEST FITTED DISTRIBUTION MODEL FOR REMOVING EXAMINERS' BIAS

S. Sahu[1], K. Harirajan[2], S. Boda[3] & S. Sahu[4]

[1]Professor, Department of Fishery Economics and Statistics, West Bengal University of Animal and Fishery Sciences, West Bengal, Kolkata, India

[2]Chairman, West Bengal Police Recruitment Board, Kolkata, West Bengal, India

[3]Research Scholar, Department of Fishery Economics and Statistics, West Bengal University of Animal and Fishery Sciences, West Bengal, Kolkata, India

[4]Department of Zoology, City College, University of Calcutta, Kolkata, West Bengal, India

ABSTRACT

The selection process of recruitment in any post poses two types of problems for an authority. The first problem is related to both the online or offline exam process, in which two or more test paper sets are used (difficulty bias) or several examiners evaluate the same test paper (examiner's bias). This necessitates normalization of the scores and the examination administrators use the equi-percentile method for this purpose. With this equi-percentile method, both the strict and lenient evaluation can be brought on a par with same scale level. The second type of problem relates to a similar constraint faced when differences exist in difficulty level for two or more sets of question papers. The objective of the present study is to statistically convert raw scores to scaled scores using equi-percentile method so as to overcome these problems. The study was conducted upon the examination scores of a subject "Bengali" over a sample size of 5525 examinees evaluated by 23 examiners during April-May 2019. The answer scripts were randomized and then placed before different examiners after proper coding. This satisfies the assumption of normality for each individual examiner. But as the evaluation process differs due to the inherent bias of each examiner, the equi-percentile method has to be applied to smoothen out the evaluation bias. In this case, examiner having maximum median (raw score) is considered as reference. So all the other marks given by different examiners are transferred to the same distribution pattern prevailed with the reference Examiner. In this case, raw scores for reference Examiner is converted to percentile rank. Then considering percentile values as independent parameter and raw marks as dependent parameters, an equation is formulated after checking the nature of the curve which fits best to the data. Then the same procedure is replicated for all the examiners. The Examiner's bias cannot be removed even if equi-percentile methodology is performed for each of the subjects for each examinee because the percentiles are not additive in nature (because it is a rank, which is a relative measure) therefore, there shall be a problem while preparing the final merit list. Now by converting all the raw scores to scaled scores by method of distribution based equalization of scores depending upon the percentile rank solves the problem, as all the raw scores converted to scaled scores are absolute values and hence based on this the merit list could be prepared.

KEYWORDS: *Equi-Percentile Method, Median, Percentile Rank, Examiners Bias, Fitted Distribution, Scaled Scores*

Article History

Received: *03 Mar 2020* **|** ***Revised:*** *06 Mar 2020* **|** ***Accepted:*** *14 Mar 2020*

INTRODUCTION

During the selection process towards recruitment to any post, the authority often comes across two types of problems. The first problem is related to both the online or offline exam process, in which two or more test paper sets are used or several examiners evaluate the same test paper. This necessitates normalization of the scores and the examination administrators use the equi-percentile method for this purpose. This implies that if an answer sheet for a particular paper set is evaluated by a lenient examiner and the answer sheet for the same paper set is evaluated by a strict examiner, then it may result in lesser marks for the examinees whose papers are evaluated by the latter. Under such a situation, the equi-percentile method is used to bring the strictly corrected answer sheet on par with the lenient one through a statistical method. In addition to several prestigious entrance exams, the IBPS or Bankers' exam, many other exams like the UP Police Recruitment & Promotion Board etc. use the equi-percentile method. With this equi-percentile method, both the strict and soft evaluation can be brought to one scale level. There is a possibility that a student claims that if his answer sheet had been assigned to a dove (lenient examiner), then his marks would have been way up the scale than he actually fared at the hands of a hawk (strict examiner). To address such a case, it is essential that equi-percentile method is used to bring both of them on par. The second type of problem relates to a similar constraint faced when difference exists in difficulty level for two or more sets of question papers. This problem can also be addressed by the cited methodology. The objective of the present study is to statistically convert raw scores to scaled scores using equi-percentile method so as to overcome the above discussed problems.

Equating is a statistical procedure commonly used in testing programs where administrations across more than one occasion and more than one examinee group can lead to over exposure of items, threatening the sanctity of the test. To the extent that behavioral measures are to be used interchangeably, the outcome scores need to be equated or made comparable. The process of deriving a function mapping score on an alternate form of a test onto the scale of the anchor form, such that after equating, any given scaled score has the same meaning regardless of which test form was administered. Equating methods can be used to adjust for differences in difficulty across alternate forms/ judgments, resulting in comparable score scales and more accurate estimates in most of the cases for different sets of examinees examined by different sets of examiners. Here, it is assumed that there exists an inherent rating bias in the evaluation of the answer scripts by different examiners. It is further assumed that an examiner is homogenous in respect of his/her rating in respect of his/her examinees but heterogeneous with other examiners. This is supported by the following literature. Statistical equating defines a functional relationship between multiple test score distributions and thereby between multiple score scales. When the test forms have been created according to the same specifications and are similar in statistical characteristics, this functional relationship is referred to as an equating function and it serves to translate scores from one scale directly to their equivalent values on another. Whether score distributions are based on samples from a single examinee population or different examinee populations (these are referred to as equating designs), if the appropriate assumptions are met the equating function can be generalized to other examinees. (Holland &Dorans, 2006). Equating with the equivalent groups design, that is, equating in their simplest and most general form, are referred to here and in the equate package as equating types. These can be categorized as either linear, including mean or linear equating, or nonlinear, equi-percentile equating. An additional nonlinear type supported in the equating package is circle-arc equating, as recently introduced by Livingston and Kim (2009). In this study, the methodology of equi-percentile equating was adopted. Percentile of a candidate will reflect how many candidates have scored below that candidate in that batch.

REVIEW OF LITERATURE

Lawton et. al (2016) have demonstrated that one can convert University of Pennsylvania Smell Identification Test (UPSIT) to Brief-SIT (B-SIT) or Sniffin' 16, and Sniffin' 12 to 16 scores in a valid way. This can facilitate direct comparison between tests aiding future collaborative analyses and evidence synthesis. They used two incident cohorts of patients with PD who were tested with either the Sniffin' 16 (n=1131) or UPSIT (n=980) and a validation dataset of 128 individuals who took both the tests used the equi-percentile and Item Response Theory (IRT) methods to equate the olfaction scales. The equi-percentile conversion suggested some bias between UPSIT and Sniffin' 16 tests across the two groups.

Brossman and Lee (2013) observed Score and True Score Equating Procedures for Multidimensional Item Response Theory. The purpose of this research was to develop observed score and true score equating procedures to be used in conjunction with the Multidimensional Item Response Theory (MIRT) framework. Three equating procedures—two observed score procedures and one true score procedure were created and described in detail. One observed score procedure was presented as a direct extension of Unidimensional IRT (UIRT) observed score equating and is referred to as the ''Full MIRT Observed Score Equating Procedure.'' The true score procedure and the second observed score procedure incorporated unidimensional approximation procedures to equate exams using UIRT equating principles. These procedures are referred to as the ''Unidimensional Approximation of MIRT True Score Equating Procedure'' and the ''Unidimensional Approximation of MIRT Observed Score Equating Procedure,'' respectively. Three exams were used to conduct UIRT observed score and true score equating, MIRT observed score and true score equating, and equi-percentile equating. The equi-percentile equating procedure was conducted for the purpose of comparison because this procedure does not explicitly violate the IRT assumption of unidimensionality. Results indicated that the MIRT equating procedures performed more similarly to the equi-percentile equating procedure than the UIRT equating procedures, presumably due to the violation of the unidimensionality assumption under the UIRT equating procedures.

Livingston and Kim (2010) in their paper entitled Random-Groups Equating with samples of 50 to 400 Test Takers employed Five methods for equating in a random groups design and were investigated in a series of resampling studies with samples of 400, 200, 100, and 50 test takers. It was done by the criterion equating, that was the direct equi-percentile equating in the group of all test takers. Equating accuracy was indicated by the root-mean-squared deviation, over 1,000 replications, of the sample equating from the criterion equating. The methods investigated were equi-percentile equating of smoothed distributions, linear equating, mean equating, symmetric circle-arc equating, and simplified circle-arc equating. Steenoven et. al (2014) in the article Conversion Between Mini-Mental State Examination, Montreal Cognitive Assessment, and Dementia Rating Scale-2 Scores in Parkinson's Disease, Mini-Mental State Examination (MMSE), Montreal Cognitive Assessment (MoCA), and Dementia Rating Scale-2 (DRS-2) have been used extensively for cognitive screening in both clinical and research settings of Parkinson's disease. The aim of this study was to apply a simple and reliable algorithm for the conversion of MoCA to MMSE scores in PD patients. A secondary aim was to apply this algorithm for the conversion of DRS-2 to both MMSE and MoCA scores. The cognitive performance of a convenience sample of 360 patients with idiopathic PD was assessed by at least two of these cognitive screening instruments. Then, it was used to develop conversion scores between the MMSE, MoCA, and DRS-2 using equi-percentile equating and log-linear smoothing.

MATERIALS AND METHODS

In classical test theory for mean equating, it simply adjusts the distribution of scores so that the mean scores of one examiner is comparable to the mean scores of the other examiner. While mean equating is advantageous because of its simplicity, it lacks flexibility, namely accounting for the possibility that the standard deviations of the scores differ. Linear equating adjusts so that the two examiners have a comparable mean and standard deviation. There are several types of linear equating that differ in the assumptions and mathematics used to estimate parameters. Equi-percentile equating determines the equating relationship as one where a score could have an equivalent percentile on either form. This relationship can be nonlinear. Unlike with Item Response Theory (IRT), equating based on classical test theory is somewhat distinct from scaling. Equating is a raw-to-raw transformation in that it estimates a raw score on Form B that is equivalent to each raw score on the base Form A. Any scaling transformation used is then applied on top of, or with, the equating.

In this case, the study was conducted upon the examination scores of subject Bengali over a sample size of 5525 examinees evaluated by 23 examiners during April-May 2019. The answer scripts were randomized and then placed before different examiners after proper coding. This satisfies the assumption of normality for each individual examiner. But as the evaluation process differs due to the inherent bias of the examiner, the equi-percentile method has to be applied to smoothen out the rating bias. Here,

- The examination has no specified and distinct guideline for awarding marks as it is vernacular and subjective. So the marks will obviously differ from examiner to examiner on the same style and content of writing.
- As the answer scripts are being examined by multiple examiners, the marking pattern will be different for different examiners, the marks scoring pattern depends upon the difficulty level of checking and it varies from one examiner to another.
- Due to this variation, the scores can be normalized using equi- percentile method to take care of the difference in difficulty of checking levels, so that no candidate feels he/she is at a loss because he/she is judged by a certain examiner.

But after converting all the raw scores to a scaled score for each examiner through equi-percentile method, it is acceptable only when a clubbed rank is made. For different examiners, all the raw scores are converted to percentile ranks corresponding to each examiner. This ensures that the hidden/underlying distribution corresponding to each examiner is smoothed out converting to a standard scale i.e., percentile scale. This methodology is appropriate for selection procedure where there is no interview and the selection solely depends upon the written exam scores. In this case all the percentile ranks corresponding to different examiners are clubbed together to prepare the selection list on the basis of percentile ranking. This methodology is being followed in the following examinations as seen recently viz. RRB, NTPC, CAT, MAT, IBPS, UPPR & PB. The drawback of the above procedure is non-incorporation of the underlying distribution pattern to the scores. The rectification can be described in the following manner. For this reason, one examiner was chosen and considered to be standard of reference. The distribution equation for that reference examiner was found by the method of multivariate analysis. In that equation, percentile rank is considered as an independent parameter and raw scores is considered as dependent parameter. Then percentile rank corresponding to each raw score of each examiner is fitted to the mentioned distribution of the reference examiner and by this way every raw mark awarded by each examiner is scaled to

this particular distribution generating the scaled scores for each individual examinee. Then by clubbing all the scaled scores of the all the examinees, the candidates are selected for the next stage of recruitment or say a Personality Test/Interview, as the case may be. But in this case along with written examination, there is an interview procedure which gives marks on an absolute scale. This problem may be solved by again making the percentile ranking for selected candidates in the case of interview. Then the written percentile ranks and interview percentile rank can be clubbed assigning some weightage to these two parameters. These weights may be the ratio of the maximum marks assigned to each test or paper. The drawback of the cited procedure can be depicted in the following ways: as the scores are converted to ranks, the weighted method will not give the desired level of efficiency to the selection procedure. The rectification can be done after completing the interviewers by all the interviewers, raw interview scores will again be converted to scaled scores by applying the previous procedure. Ultimately in the case of final selection, as all the scores where there is a possibility of evaluators' bias is thus removed by the above procedure of equi-percentile equating method fitted to some reference distribution generating the scaled scores on an absolute scale. These scaled scores can be used for selection purpose compatible to other raw scores which are free from human bias.

The collected data were statistically analyzed through SPSS 21.0 and Microsoft Excel Work sheet.

RESULT AND DISCUSSIONS

The study was conducted upon the examination scores of one subject viz, 'Bengali' over a sample size of 5525 examinees evaluated i.e. distributed among 23 examiners during April-May 2019. The answer scripts were randomized and then placed before different examiners after proper coding. This satisfies the assumption of normality for each individual examiner. But as the evaluation process differs due to the inherent bias of the examiner commonly known as the rating bias the equi-percentile method has to be applied to smoothen out the rating bias. To judge about the central tendency of each examiner, the following table depicts the descriptive statistics for the selected sample.

Table 1: Distribution of Marks (Bengali) and their Descriptive Statistics (Arranged on the Basis of Median

Code	Examiner	Freq	Mean	SD	Median	Min	Max	Code	Examiner	Freq	Mean	SD	Median	Min	Max
BO	EX-15	200	15.05	6.53	16	0	31	BP	EX-16	300	22.08	6.10	22	7	37
BT	EX-20	300	14.86	5.33	16	0	28	BA	EX-1	100	23.38	4.91	23	7	35
BQ	EX-17	300	17.73	6.80	17	0	35	BL	EX-12	300	22.89	3.57	24	0	31
BR	EX-18	300	16.09	4.60	17	2	31	BB	EX-2	100	24.82	3.11	25	13	31
BS	EX-19	300	16.99	4.54	17	0	30	BJ	EX-10	300	23.77	6.66	25	0	36
BE	EX-5	100	20.15	3.58	20	8	30	BV	EX-22	300	24.00	6.52	25	4	37
BI	EX-9	300	18.96	6.65	20	0	36	BC	EX-3	100	25.11	2.43	26	15	29
BU	EX-21	300	19.87	6.61	20	3	35	BG	EX-7	100	25.12	4.30	26	5	33
BN	EX-14	300	20.33	5.41	21	5	36	BH	EX-8	300	25.37	4.83	26	0	39
BD	EX-4	100	22.19	3.72	22	15	31	BW	EX-23	325	25.11	4.80	26	6	35
BK	EX-11	400	21.27	3.95	22	0	30	BF	**EX-6**	100	27.49	6.37	**28**	2	39
BM	EX-13	300	20.98	5.19	22	0	31	**Total Frequency= 5525**							

From the table 1, it is clear that maximum median is 28 corresponding to Examiner-6 i.e. code BF. Statistically, a median of medians is the thumb-rule. But any of the examiners could be chosen as reference (examiner). This also satisfies the method since the underlying distribution of the reference examiner is considered. However, taking the median of medians as the reference examiner may lead to examinees with higher raw scores being awarded lower scaled scores

resulting in grievances for these test takers. In this case, maximum median is 28 corresponding to Examiner-6 i.e. BF. So all the other marks given by different examiners are transferred to the same distribution that prevailed with Examiner -6. The following Histogram, Box Plots and Normal Q-Q Plots reveals the nature of the data for further analysis.

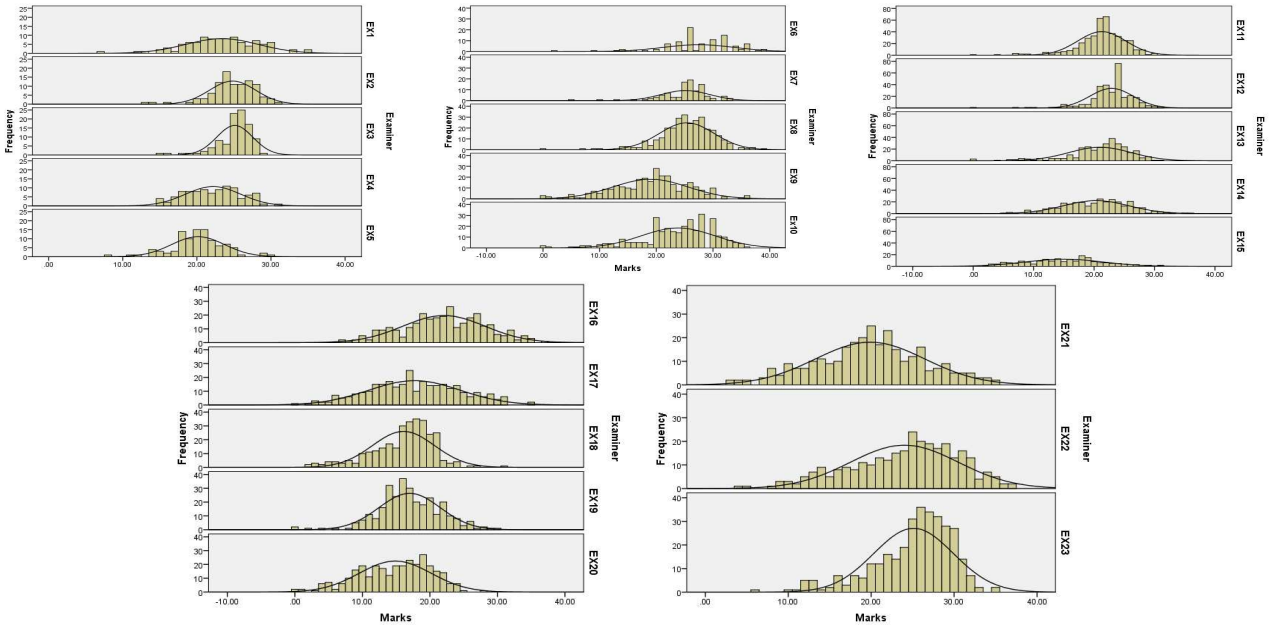
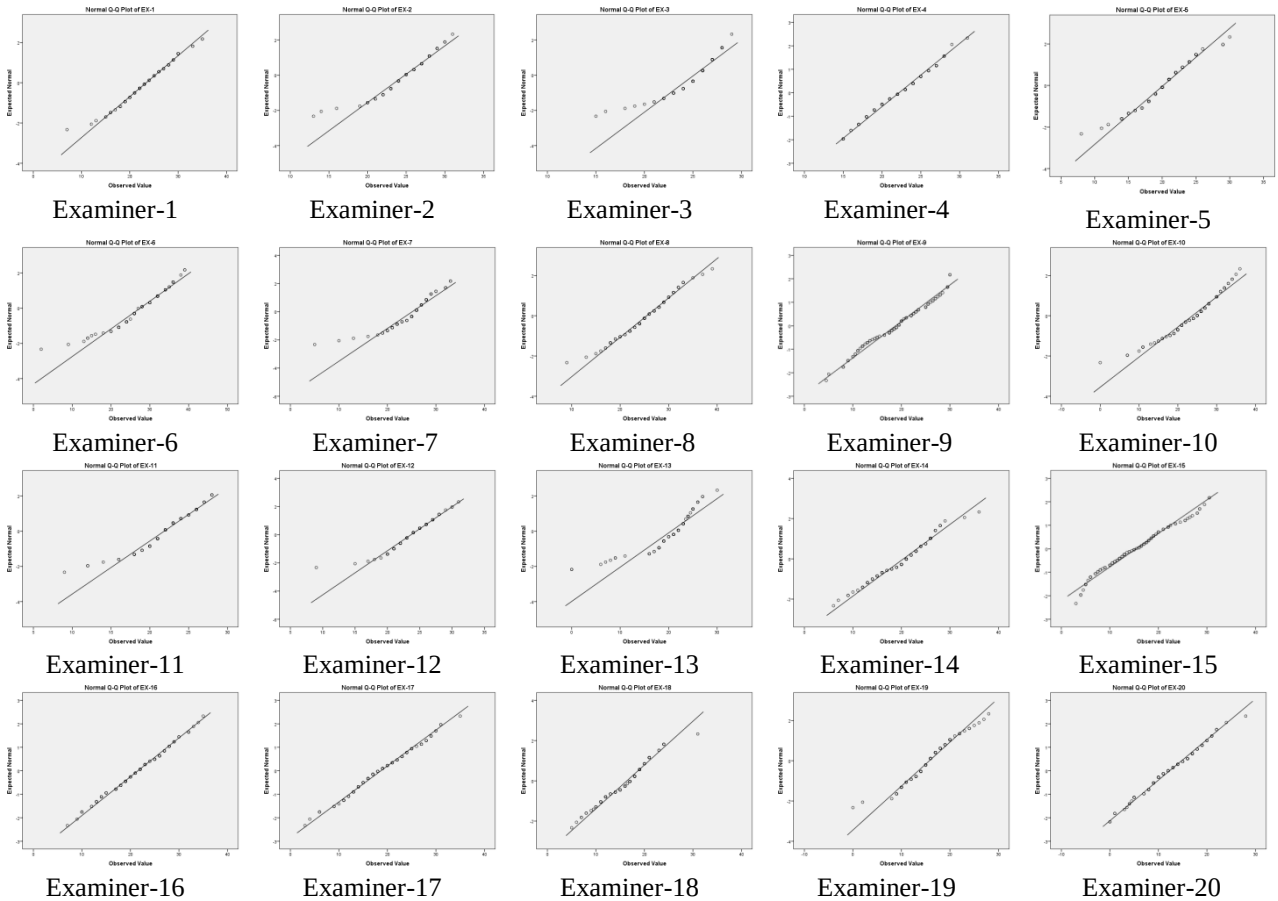
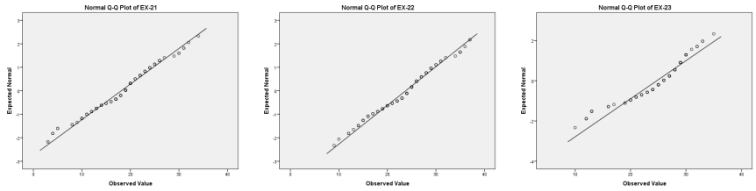


Figure 1: Frequency Distribution of Marks given by 23 Examiners.





Examiner-21 Examiner-22 Examiner-23

Figure 2: Normal Q-Q Plots of Marks given by 23 Examiners.

Percentile Ranking Conversion Table

In the case of marks in Bengali, raw score for Examiner -6 (taken as reference) is scaled to percentile rank. Then considering percentile as independent parameter and raw marks as dependent parameters an equation is formulated after checking the nature of the curve which fitted best to the data. For Examiner-6: independent parameter: percentile, dependent parameter: raw marks.

$$\text{Percentile rank} = \frac{\text{Number of candidates appeared in examination} - \text{Rank of candidate}}{\text{Number of candidates appeared in examination} - 1} \times 100$$

Model Summary and Parameter Estimates

Dependent Variable: Marks Independent variable, Percentile

Table 2: Results of Curve Estimation

Sl no	Equation	R^2	Sig.
1.	Linear	.885	.000
2.	Quadratic	.893	.000
3.	**Cubic**	**.952**	**.000**
4.	Compound	.562	.000
5.	Growth	.562	.000
6.	Exponential	.562	.000

The independent variable (percentile) contains non-positive values. The minimum value is.00. The Logarithmic and Power models cannot be calculated. The independent variable (percentile) contains values of zero. The Inverse and S models cannot be calculated.

From the above table The R^2 value is maximum for the case of Cubic equation. So Cubic equation will explain more or less 95.2 % of the variability at 1% significant level. So, it is evidently clear that for cubic equation the data fitted best. This is also supported by the following different fitted curve.

Table 3: Coefficients for the Multivariate Regression Curve Assuming a Cubic Model

Sl No		B (Coefficients)	Sig.
1.	percentile	0.781	0.00000
2.	percentile ** 2	-0.013	0.00000
3.	percentile ** 3	0.000082	0.00000
4.	(Constant)	11.086	0.00000

The regression equation to be obtained for examiner 6 is:

$$Y(\text{Scaled marks}) = 11.086 + (0.781 \times \text{percentile rank}) + (-0.013 \times \text{percentile rank}^2) + (0.000082 \times \text{percentile rank}^3)$$

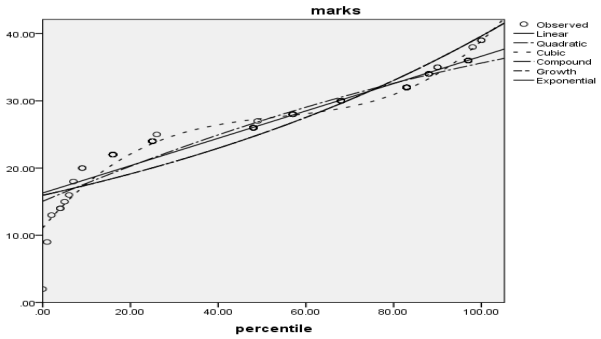


Figure 3: Checking the Best Fitted Curve.

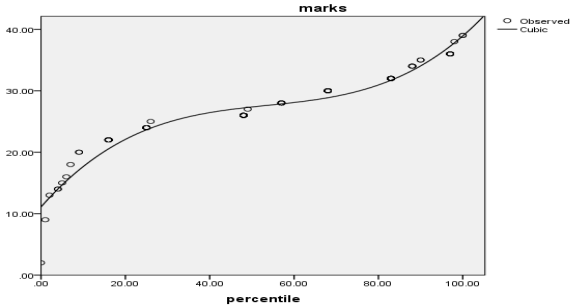


Figure 4: Cubic Curve between Independent and Dependent Variables.

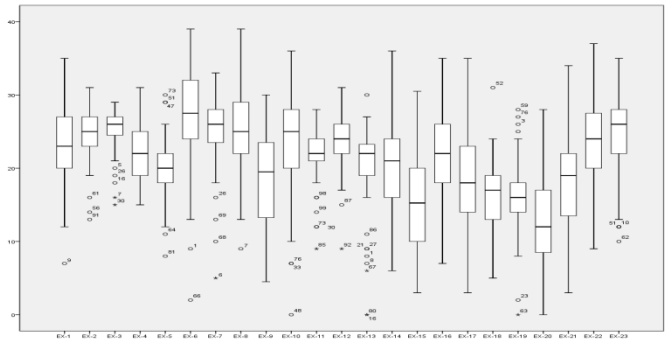


Figure 5: Box-Plots of Raw Marks given by 23 Examiners.

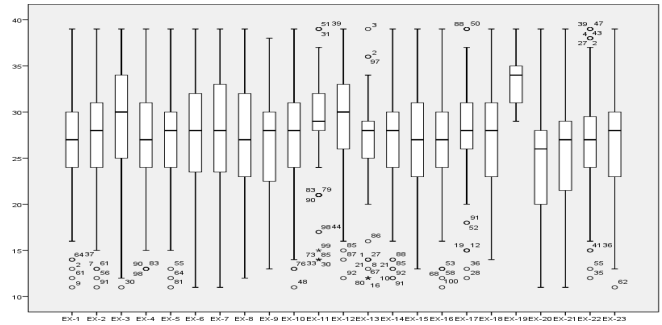


Figure 6: Box-Plots of Scaled Scores given by 23 Examiners.

Table 4: Table Showing Some Raw and Scaled Score from the Raw Data

Code	EX-6 Raw Marks	Percentile Rank	Scaled Score	Code	EX-6 Raw Marks	Percentile Rank	Scaled Score
BF1	9	1	12	BF7	32	83	32
BF2	14	4	14	BF8	36	97	37
BF3	39	100	39	BF9	34	88	33
BF4	39	100	39	BF10	32	83	32
BF5	32	83	32	BF11	36	97	37
BF6	30	68	29				

CONCLUSIONS

An equi-percentile method would be statistically rigorous if only there was a single subject paper of a descriptive type which needs to be scaled for various examiners to even out rating bias. Similarly, an objective type test paper administered over several sessions, that is, several test papers of different difficulty levels would also call for some method to even out the differences in the test papers.

However, in case of different test papers carrying different maximum marks such as for admission, recruitment or academic tests where marks are awarded for test papers on different subjects, case studies, group discussion, interview or personality tests a simple equi-percentile method would not suffice. In all such cases, the underlying distribution of marks awarded by different examiners transferred to the distribution of the reference examiner would still place the raw scores in a percentile rank. A very important point that may be missed out is that there is a need for the scaled marks (absolute value) of different subject papers to enable adding them up for preparing a merit list of test takers. An equi-percentile method followed by a study of the underlying distribution of any one examiner (reference examiner) and transferring all the raw scores awarded by all the other examiners by converting them to a percentile rank with reference to each examiner for all his examinees and to a scaled score by using the reference examiner's distribution. This would give scaled scores (absolute values) which would enable the examination administrators to prepare a merit list. This is however not possible in the case of a simple equi-percentile method.

REFERENCES

1. *Lawton M, Hu MTM, Baig F, Ruffmann C, Barron E, Swallow DMA, Malek N, Grosset KA, Bajaj NP, Barker RA, Williams N, Burn DJ, Foltynie T, Morris HR, Wood NW, May MT, Grosset DG, Ben-Shlomo Y, Equating scores of the University of Pennsylvania Smell Identification Test and Sniffin' Sticks test in patients with Parkinson's disease, Parkinsonism and Related Disorders (2016), doi: 10.1016/j.parkreldis.2016.09.023.*
2. *Livingston, S. A., & Kim, S. (2010). Random□groups equating with samples of 50 to 400 test takers. Journal of Educational Measurement, 47(2), 175-185.*
3. *Brossman, B. G., & Lee, W. C. (2013). Observed score and true score equating procedures for multidimensional item response theory. Applied Psychological Measurement, 37(6), 460-481.*
4. *Van Steenoven, I., Aarsland, D., Hurtig, H., Chen□Plotkin, A., Duda, J. E., Rick, J, & Moberg, P. J. (2014). Conversion between mini□mental state examination, montreal cognitive assessment, and dementia rating scale□2 scores in Parkinson's disease. Movement Disorders, 29(14), 1809-1815.*
5. *Holland, P.W. & Dorans, N.J.(2006). Linking and Equating. In R.L. Brennan (Ed.), Educational measurement (4th. Ed.,pp 187-220). Westport, CT: American Council for Education/ Prager.*

AUTHOR PROFILE

Professor Somen Sahu, a Statistician and Fishery Economist by profession, is a Professor and Head of Fishery Economics & Statistics, Faculty of Fishery Sciences, West Bengal University of Animal And Fishery Sciences, Kolkata. Prof. Sahu completed B.Sc. Honours in Statistics from Ramkrishna Mission Residential College, Narendrapur (Calcutta University) in 1991, Post-graduated in Statistics from Burdwan University in 1993 (1st. class 2nd.), M.B.A. (P.G.D.P.M.) from National Institute of Personnel Management in 1995 with Gold Medal. He completed his Ph.D. from Jadavpur University in 2006. He was a National Scholar. His areas of interest are Bio-Statistics, Statistical Software Handling, Biomonitoring, Management Information System and Extension Education in different Agricultural fields. He introduced a new Model viz. “Dr. Sahu’s Networking Model” which is an econometric model for optimization of production-technology (enhancement of Productivity Index) in inland fisheries which was adopted by Department of Fisheries, Government of West Bengal. He is a field based fishery research scientist and academician of repute. Another pioneer work in his credit that he has successfully introduced the concept of deep sea marine cage culture which is acclaimed by all concerned relating to fishery sciences. He published more than 40 (Forty) research publications in different peer reviewed National and International journals, guided and guiding more than 61 (Sixty One) [Master's (42) and Ph.D.(19)] scholars. He had also bagged the Intellectual Property Right accreditation applicable over 177 countries on the globe, on the “PROF. SAHU’S METHODOLOGY OF DISTRIBUTION DEPENDENT EQUALIZATION OF SCORES TO REMOVE EXAMINER’S BIAS AND/OR DIFFICULTY BIAS” as a co-author. He is the founder Secretary of International Organisation of Biological Data Handlers. He has life membership with various scientific & professional societies & organizations. At present, he is associated with different organizations in State and National level. For his active contributions in the field, he has been awarded "LIFE TIME ACHIEVEMENT AWARD-2017", "MATSYA SATHI SAMMAN-2018", “BEST SCIENTIST AWARD-2020” during Bengal Aqua Expo (over different years) by Sri Chandranath Sinha, Hon'ble MIC, Dept. of Fisheries, Govt. of W.B.. He received the “BEST TEACHER AWARD-2019”, F.F.Sc., from his University. He is also associated with the following organizations in different capacity viz. MARINE ADVISOR, Department of Fisheries, Govt. of West Bengal, Expert Member of ICAR-CMFRI (Central Marine Fisheries Research Institute), Govt. of India, Expert Member of ESSO-INCOIS (Indian National Center for Ocean Information Services), Govt. of India, Honorary Statistical Adviser, West Bengal Police Recruitment Board, Govt. of West Bengal, Editor In-Chief, The Indian Journal of Agriculture Business, Co-Principal investigator, All India Network Project on Mariculture, ICAR.