International Journal on Emerging Technologies 11(2): 295-303(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Equalization of Scores through Statistical Modeling towards Removing Examiners' Bias for Arithmetic and Reasoning Paper

S. Sahu[1], K. Harirajan[2], G. Mahapatra[3] and S. Sahu[4]

[1]*Professor and Head, Department of Fishery Economics and Statistics, West Bengal University of Animal and Fishery Sciences, Kolkata (West Bengal), India.*

[2]*Chairman, West Bengal Police Recruitment Board, Kolkata (West Bengal), India.*

[3]*Associate Professor, Department of Computer Science, Asutosh College, University of Calcutta, Kolkata, West Bengal, India and Research Fellow, Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Off-Campus Deoghar (Jharkhand), India.*

[4]*Department of Zoology, City College, Calcutta University, Kolkata (West Bengal), India.*

(Corresponding author: G. Mahapatra)

(Received 28 December 2019, Revised 04 March 2020, Accepted 16 March 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Making rational sorted list of candidates after evaluation of subjective test evaluated manually by the strict, lenient and moderate etc. different category of examiners is the practical challenges to face by the recruitment boards and others. Equi-percentile based normalization is a very common process of equalization of scores and used by the different popular recruitment authorities to cope such anomalies. But, if there is more than one subject, this methodology fails because of the inapplicability of the additive property towards rank values. The same constraint is also faced when there are differences in difficulty level for two or more sets of question papers for a single recruitment due to the size of the examinees. Present study is proposing a statistical model to remove the existence of examiners' bias in the evaluation of the multi-subjects and multi-level difficulties based on individual ranking process. Effectiveness and viability of the proposed model has been successfully tested with practical data sets.

Keywords: Equalization of scores, Equi-percentile method, Examiners' bias, Median, Statistical distribution.

Abbreviations: UPSIT, University of Pennsylvania Smell Identification Test; IRT, Item Response Theory; MIRT, Multidimensional Item Response Theory; Q-Q plot, Quantile-Quantile plot; UIRT, Uni-dimensional IRT; MoCA, Montreal Cognitive Assessment; MMSE, Mini-Mental State Examination.

I. INTRODUCTION

Preparing ordered selection list of the candidates in recruitment process nowadays involves primarily either pen-paper or computer based examinations recommending authorities and/or different boards. This method has drawbacks because, due to huge volume, the answer scripts are evaluated manually by several examiners who may be strict, lenient or moderate in their evaluation. This result in rating bias for the scores and hence the evaluations cannot be used directly in the preparation of the selection list. To remove such bias in the rating system the application of equi-percentile method is very common technique. In this case, the examination administrators use the equi-percentile method for normalization. With this equi-percentile method, a toughly evaluated paper and a softly evaluated one are brought to one scale level. This paper presents a case study of a similar situation where the examination scores of Arithmetic and Reasoning paper (subjective in nature) of 1238 examinees were evaluated by eight examiners (after proper coding) during January- 2020 (for a Graduate level entry post) were statistically converted to a scaled score using equi-percentile method and adopting Statistical Modeling approach. In this case maximum median is 33 corresponding to Examiner-46 i.e. EX-46. So, all other marks given by different examiners are transferred to the same distribution which prevailed with Examiner-46. Statistically, a median of medians is the thumb-rule and any of the examiners could be chosen as a reference (examiner). This also satisfies the method since it would be considering the underlying distribution of the reference examiner. 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 the test takers. Thus all the raw scores converted to scaled scores. The drawback of the previous study/method adopted by several recruitment authority is that the Examiners' Bias/Difficulty Bias (as the case may be) cannot be removed even if equi-percentile methodology is performed for each of the subjects (sets of questions) for each examinee because the percentiles are not additive in nature (because it's a rank, which is a relative measure). This leads to a problem while preparing a merit list.

II. EARLIER WORKS

To mention the actual difficulties and practical situation here is a statement that has been considered in the per view of this study "Staff Selection Commission has been conducting various examinations in multiple batches because of large number of candidates and difficulties in getting adequate educational institutions for holding the examinations in a single batch. For perhaps the first time in its history, the number of applicants in a single examination exceeded one million when the Combined

Higher Secondary Level Examination, 2010 for the recruitment of Lower Division Clerks and Data Entry Operators, elicited response from over 16 lakh candidates, with approx. 21% of them applying online. This would require the Examination, rescheduled on 27 & 28.11.2010 (in view of Common Wealth Games), to be held in at least three batches. The Commission, with the help of experts, has striven to construct question papers of comparable difficulty level. While such an exercise is theoretically possible, in practice it is impossible to have two or more question papers of identical difficulty levels. Even if the difficulty levels of question papers vary slightly, candidates taking more difficult papers may be at a disadvantage viz-a-vis others. Therefore, there is a need for equating of the marks in examinations involving multiple batches and question papers… The Commission had examined the views of an Expert Group, constituted by it with the approval of Government of India in 2009, on this issue. The Commission had placed before the Expert Group that the technique to be followed for equating should be transparent, easily comprehensible to the candidates, acceptable to experts and prove itself in Courts of Law if and when challenged. This was accepted by the Expert Group which further advised the Commission to place a paper on the technique on its website for adequate time, give publicity to such placement through the media, invite comments, observations and suggestions and decide on adopting the technique thereafter Equating is a statistical process that is used to adjust scores on multiple question papers so that scores on the forms can be used interchangeably. It adjusts for differences in difficulty among Question Papers that are built to be similar in difficulty and content. As per the report the expert committee viewed about four methods of Equating viz. (i) Median/Mean Equating, (ii) Linear Equating (Based on mean and S.D.), (iii) Equi-percentile Equating, (iv) Equating using Item Response Theory. Among these methods, SSC proposes to use the Equi-percentile Method in view of its simplicity" [1]. In one study Lawton *et al*., (2016) demonstrated the method to 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 facilitated direct comparison between tests aiding future collaborative analyses and evidence synthesis [2]. Lawton *et al*., (2016) used the equi-percentile and Item Response Theory (IRT) methods to equate the olfaction scales and validated dataset of 128 individuals who took both tests, the Sniffin' 16 (n=1131) or UPSIT (n=980) [2]. The equi-percentile conversion suggested some bias between UPSIT and Sniffin' 16 tests across the two groups. The IRT method shows very good characteristics between the true and converted Sniffin' 16 (delta mean = 0.14, median = 0) based on UPSIT. The equi-percentile conversion between the Sniffin' 12 and 16 item worked well (delta mean = 0.01, median = 0). Lee (2013) develop observed score and true score equating procedures to be used in conjunction with the Multidimensional Item Response Theory (MIRT) framework [3]. Three equating procedures—two observed score procedures and one true score procedure—were created. One observed score procedure was presented as a direct extension of Uni-dimensional 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 uni-dimensional approximation procedures to equate exams using UIRT equating principles. These procedures are referred to as the "Uni-dimensional Approximation of MIRT True Score Equating Procedure" and the "Uni-dimensional 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 uni-dimensionality. 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 uni-dimensionality assumption under the UIRT equating procedures. Livingston and Kim (2010) proposed five methods for equating in a random groups design with samples of 50 to 400 Test Takers. The criterion equating was the direct equi-percentile equating in the group of all test takers [4]. 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. The circle-arc methods produced the most accurate results for all sample sizes investigated, particularly in the upper half of the score distribution. The difference in equating accuracy between the two circle-arc methods was negligible; van Steenoven *et al.,* [6] applied a simple and reliable algorithm for the conversion of Montreal Cognitive Assessment (MoCA) to Mini-Mental State Examination (MMSE) scores in PD patients. Further, the same algorithm was applied for conversion of Dementia Rating Scale-2 (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. He then developed conversion scores between the MMSE, MoCA, and DRS-2 using equi-percentile equating and log-linear smoothing. The conversion score tables reported enable direct and easy comparison of three routinely used cognitive screening assessments in PD patients. The classical test theory for mean equating adjusts the distribution of scores so that the mean scores of one examiner are comparable to the mean scores of another. However, this method lacks flexibility, as there exists the possibility for difference in the standard deviations of the scores. Linear equating resolves this issue and adjusts in a way that the two examiners have a comparable mean and standard deviation. Based on assumptions and mathematics used, linear equating is of several types. 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. Equating is explained as transformation on raw-to-raw basis. It involves estimating a raw score on Form Y equivalent to the raw score in base form X with further application of scaling transformations. The basic research gap is that no such methodology derived to convert raw scores to scaled scores in absolute value considering the basic distribution pattern except the recent development -

"Prof. Sahu's methodology of distribution dependent equalization of scores to remove examiner's bias and/or difficulty bias" [8].

III. MATERIALS AND METHODS

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. According to Holland and Dorans (2006) [7], 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. Equating methods can be used to adjust for differences in difficulty across alternate forms/ judgments, resulting in comparable score scales and more accurate estimates of ability in most of the cases for different sets of examinees examined by different sets of examiners. Here it is assumed that there exists rating biases 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. Equating types can be categorized as either linear, including mean or linear equating, or nonlinear, equi-percentile equating. An additional nonlinear type is circle-arc equating, as recently introduced by Livingston and Kim [4]. For the present study the methodology of equi-percentile equating is adopted. The percentile of a candidate will reflect how many candidates have scored below that candidate in that batch.

The procedure of Equi-percentile equating method is discussed briefly:Informally it is used to equate scores on two tests so that the scores reflect the same percentiles should be based on same set of respondents, but often based on randomly equivalent groups. Formally, X-score x and Y-score y are linked in T if FT(x) = GT(y) When these two CDF's are continuous and strictly increasing, then this equation can always be satisfied. This is a very effective method for equating. Equi-percentile equating defines a nonlinear relationship between score scales by setting equal the percentile ranks for each score point. Specifically, the equi-percentile equivalent of a form-X score on the Y scale is calculated by finding the percentile rank in X of score i, and then the form-Y score associated with that form-Y percentile rank:

$$\mathrm{Y}(\mathrm{X_i}) = \mathrm{Q} - 1[\mathrm{P}(\mathrm{X_i})]$$

Here, $\mathrm{P}(\mathrm{X})$ is the percentile rank function in X and $\mathrm{Q} - 1[\mathrm{P}(\mathrm{X_i})$ is the inverse percentile rank function in Y. According to Kolen & Brennan (2014) [6] the process is complicated by the fact that scores are discrete and must be made continuous. Because it involves estimation at each score point, equi-percentile equating is especially susceptible to random sampling error. Smoothing methods are typically used to reduce irregularities in either the score distributions or the equating function itself.

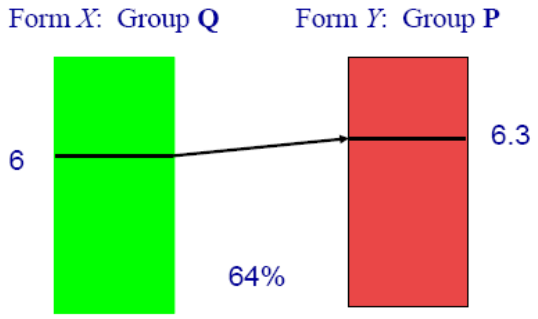


Fig. 1. Equating Method.

The study was conducted for the test of Arithmetic and Reasoning (subjective in nature) which has no specified and distinct guideline for giving marks as it is subjective. As a consequence a difference in evaluation is obvious among the examiners resulting in variation of the marks for the same style and content of writing. The marks scoring pattern for the answer sheets depends on the difficulty level of checking by the examiners and varies among the different examiners entrusted for the purpose. Such variations in scores necessitate normalization. Equi-percentile Method takes care of the difference in difficulty of checking level and resultant rating bias of the examinees. In this case by using equi-percentile method all the raw scores were first converted to a scaled score for each examiner followed by clubbed ranking. This ensures smoothing out the hidden/underlying distribution corresponding to each examiner converting it to a standard scale i.e., percentile scale. This methodology is appropriate for selection procedure where there are no further score tests or interview and the selection solely depends upon the written exam scores of only one paper or a subject. This methodology is being followed in the following examinations as seen recently viz. RRB, NTPC, CAT, MAT, IBPS, UPPR-PB and SSC. However, the above procedure has a drawback. In this method it is not possible to incorporate the underlying distribution pattern to the scores and as the ranks are not additive in nature, it cannot be used for more than one subject. Now, to rectify the problem one examiner is considered to be the standard and chosen as reference. Then the distribution equation for that reference examiner is determined by the method of multivariate analysis. In that equation percentile rank is considered as independent parameter and raw scores is considered as dependent parameter. Then percentile rank corresponding to each raw scores of each examiner is fitted to the mentioned distribution of the reference examiner and by this way every raw marks awarded by each examiner will be 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 it is possible to select the candidates for the next stage of recruitment or say a Interview/Personality Test or Viva voce, as the case may be. Moreover, in some selection procedures, a written examination is followed by an interview, 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. But as the scores are converted to ranks the weighted method will not give the desired level of efficiency to the

selection procedure. The only rectification method is that, after completing the interview by all the interviewers, scores will again be converted to scaled scores by applying the previous procedure. As all the scores where there is a possibility of evaluators' bias 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 taken for selection purpose compatible to other raw scores which are free from human bias.

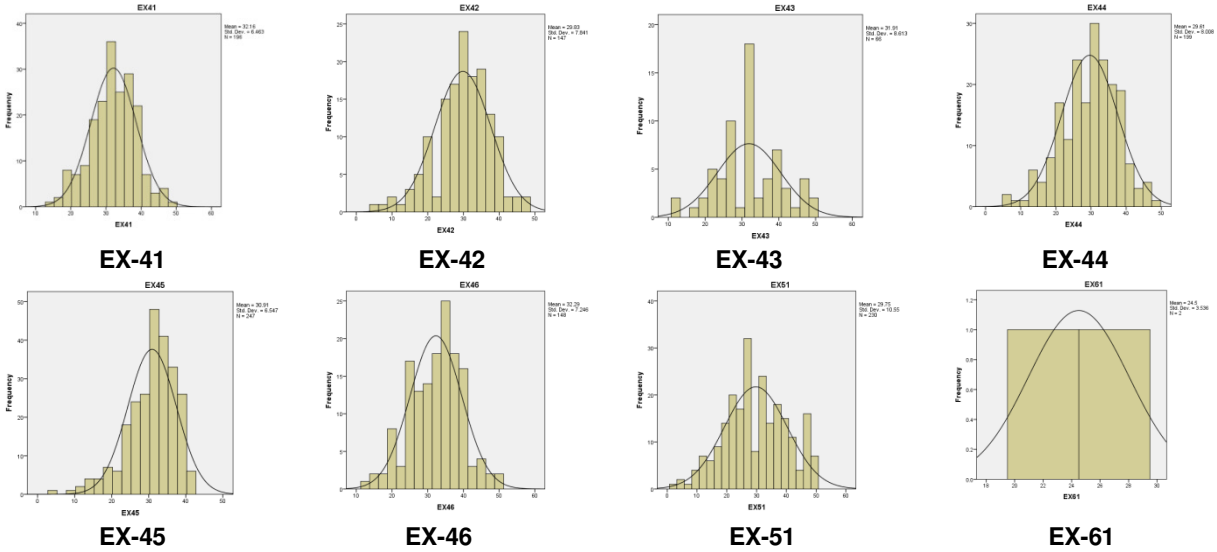
The collected data was statistically analyzed through SPSS 21.0 and Microsoft Excel Work sheet.

IV. RESULTS AND DISCUSSION

This section may each be divided by subheadings or may further divide into next heads as shown below. The present study involves the examination scores of Arithmetic and Reasoning of a sample size of 1238 examinees. The answer scripts were randomized and distributed among eight examiners for evaluation during January-2020. Although the randomized distribution satisfies the normality for each individual examiner but the inherent bias of the examiners commonly called rating bias is a major drawback. Therefore, the equi-percentile method has to be applied to smoothen out the rating bias. To judge about the central tendency of each examiner the descriptive statistics for the selected samples are shown in the Table 1.

Table 1: Distribution of Marks (Arithmetic and Reasoning) and their descriptive Statistics.

Examiner Code	N	Mean	Median	Mode	Std. Deviation	Skewness	Kurtosis	Min	Max	Percentiles		
										25 Q1	50 Q2	75 Q3
EX61	2	24.50	24.50	22a	3.536			22	27	22.00	24.50	—
EX51	230	29.75	29.00	31	10.550	-0.061	-0.536	2	50	23.00	29.00	37.00
EX44	199	29.61	30.50	32a	8.008	-0.375	0.092	6	48	24.00	30.50	35.00
EX42	147	29.83	31.00	30a	7.841	-0.586	.587	5	48	25.00	31.00	35.00
EX43	66	31.91	31.50	31	8.613	0.021	-.183	12	49	26.00	31.50	38.63
EX41	196	32.16	32.00	37	6.463	-0.247	.036	14	48	28.00	32.00	37.00
EX45	247	30.91	32.00	31	6.547	-1.015	1.361	4	42	27.00	32.00	36.00
EX46	148	32.29	33.00	34	7.246	-0.252	-0.052	13	49	27.00	33.00	37.00
TOTAL	1235											
The highest Median corresponding to EX46 is taken into consideration for deriving the required distribution.												



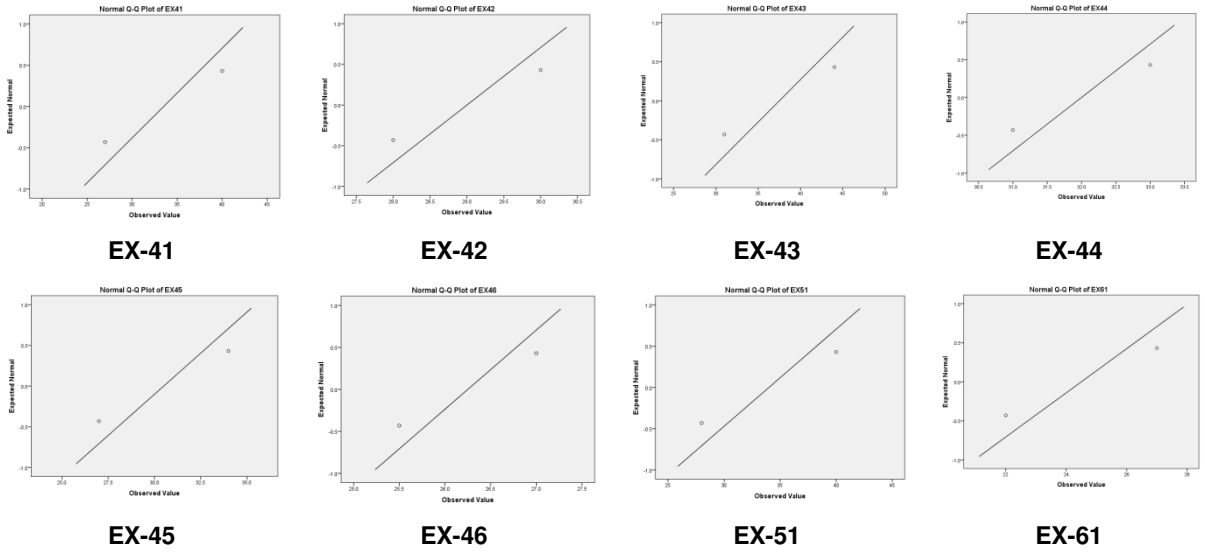
EX-41 **EX-42** **EX-43** **EX-44**

EX-45 **EX-46** **EX-51** **EX-61**

Fig. 2. Histogram with Normal curve of marks given by 8 examiners.

From the Table 1, it is evident that Maximum median is 33 corresponding to Examiner-46. Statistically, a median of medians is the thumb-rule and any of the examiners could be chosen as reference (examiner). This also satisfies the method since it would be considering the underlying distribution of the reference examiner. 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 the test takers.

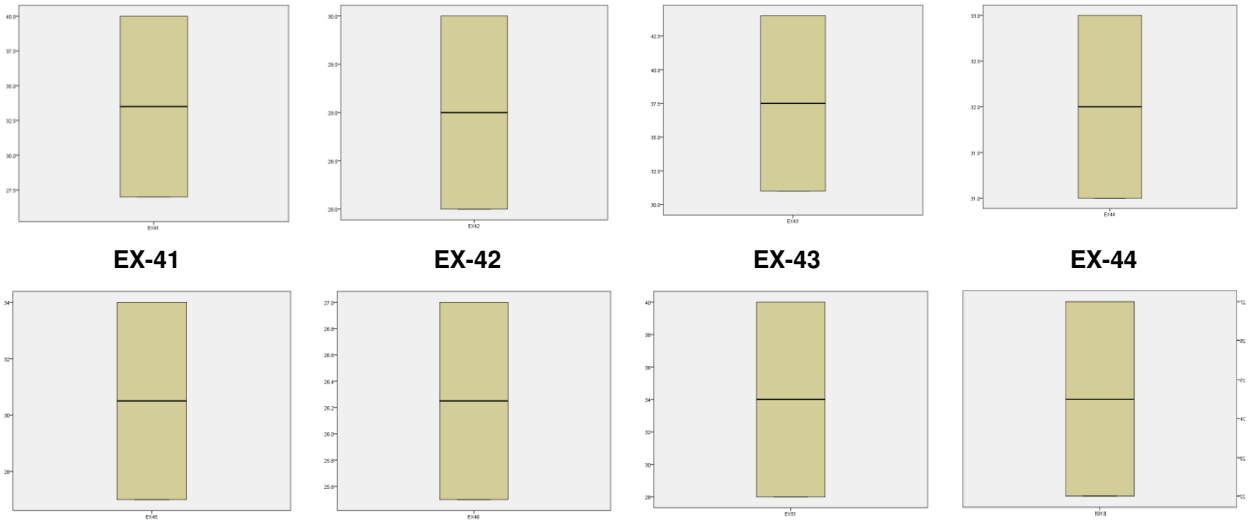
In this case Maximum median is 33 corresponding to Examiner-46 i.e. EX-46. So, all the other marks given by different examiners are transferred to the same distribution which prevailed in examiner-46. The Histogram, Box Plots and Normal Q-Q Plots shown in Fig. 2, 3, 7 and 8 are respectively revealing the nature of thedata for further analysis. Table 2 is representing the summery of the models and respective parameters, from which the R2-value is maximum in the case of Cubic equation.



EX-41 **EX-42** **EX-43** **EX-44**

EX-45 **EX-46** **EX-51** **EX-61**

Fig. 3: Normal Q-Q plots of marks given by 8 examiners



EX-41 **EX-42** **EX-43** **EX-44**

EX-45 **EX-46** **EX-51** **EX-61**

Fig. 4: Box- plots of marks given by 8 examiners

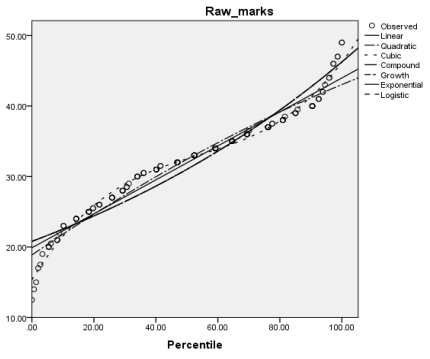


Fig. 5. Checking the best fitted curve

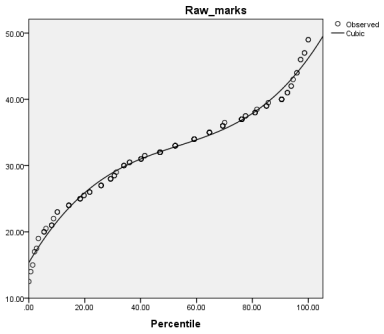


Fig. 6. Cubic curve between independent and dependent variables

Table 2: Model Summary and Parameter Estimates.

Dependent Variable: Raw_marks		
Equation	Model Summary	
	R Square	Sig.
Linear	0.951	0.000
Logarithmic[a]	—	—
Inverse[b]	—	—
Quadratic	0.954	0.000
Cubic	0.988	0.000
Compound	0.890	0.000
Power[a]	—	—
S[b]	—	—
Growth	0.890	0.000
Exponential	0.890	0.000
Logistic	0.890	0.000
The independent variable is Percentile.		
[a]The independent variable (Percentile) contains non-positive values. The minimum value is 0.00. The Logarithmic and Power models cannot be calculated. [b]The independent variable (Percentile) contains values of zero. The Inverse and S models cannot be calculated.		

Table 3: AVOVA Table for the parameters towards reference examiner (Ex-46).

	Sum of Squares	df	Mean Square	F	Sig.
Regression	7628.978	3	2542.993	4092.135	0.000
Residual	89.487	144	0.621		
Total	7718.465	147			
The independent variable is Percentile.					

Table 4: Model of raw marks and final scale score of Arithmetic and Reasoning scores (Only the first fifty examinees of Examiner 46 are depicted).

Code of Examinees	Raw marks	Rank	Percentile Rank	Scaled score	Code of Examinees	Raw marks	Rank	Percentile Rank	Scaled score
CANEX2 0239	49	1	100	45.256	CANEX2 0745	39	23	85.03401361	38.74394257
CANEX2 1022	49	1	100	45.256	CANEX2 1220	39	23	85.03401361	38.74394257
CANEX2 0265	47	3	98.63945578	44.51001888	CANEX2 0269	38.5	28	81.63265306	37.71124314
CANEX2 1040	47	3	98.63945578	44.51001888	CANEX2 0240	38	29	80.95238095	37.52098218
CANEX2 0266	46	5	97.27891156	43.79792008	CANEX2 0268	38	29	80.95238095	37.52098218
CANEX2 1008	46	5	97.27891156	43.79792008	CANEX2 0306	38	29	80.95238095	37.52098218
CANEX2 1021	44	7	95.91836735	43.11867457	CANEX2 1003	38	29	80.95238095	37.52098218
CANEX2 1219	44	7	95.91836735	43.11867457	CANEX2 1018	38	29	80.95238095	37.52098218
CANEX2 0304	43	9	94.55782313	42.47125328	CANEX2 0296	37.5	34	77.55102041	36.64399735
CANEX2 0751	42	10	93.87755102	42.15915515	CANEX2 1006	37.5	34	77.55102041	36.64399735
CANEX2 0756	42	10	93.87755102	42.15915515	CANEX2 0233	37	36	76.19047619	36.32548494
CANEX2 0255	41	12	92.5170068	41.55754074	CANEX2 0238	37	36	76.19047619	36.32548494
CANEX2 0258	41	12	92.5170068	41.55754074	CANEX2 0262	37	36	76.19047619	36.32548494
CANEX2 0279	41	12	92.5170068	41.55754074	CANEX2 0264	37	36	76.19047619	36.32548494
CANEX2 0254	40	15	90.47619048	40.70964431	CANEX2	37	36	76.19047619	36.32548494

					0271				
CANEX2 0257	40	15	90.47619048	40.70964431	CANEX2 0280	37	36	76.19047619	36.32548494
CANEX2 0290	40	15	90.47619048	40.70964431	CANEX2 0757	37	36	76.19047619	36.32548494
CANEX2 1015	40	15	90.47619048	40.70964431	CANEX2 0765	37	36	76.19047619	36.32548494
CANEX2 1020	40	15	90.47619048	40.70964431	CANEX2 1009	37	36	76.19047619	36.32548494
CANEX2 1215	40	15	90.47619048	40.70964431	CANEX2 0305	36.5	45	70.06802721	35.08649693
CANEX2 1238	40	15	90.47619048	40.70964431	CANEX2 0228	36	46	69.3877551	34.96588704
CANEX2 0281	39.5	22	85.71428571	38.96766181	CANEX2 0248	36	46	69.3877551	34.96588704
CANEX2 0227	39	23	85.03401361	38.74394257	CANEX2 0291	36	46	69.3877551	34.96588704
CANEX2 0237	39	23	85.03401361	38.74394257	CANEX2 0752	36	46	69.3877551	34.96588704
CANEX2 0259	39	23	85.03401361	38.74394257	CANEX2 1011	36	46	69.3877551	34.96588704

So, Cubic equation will explain more or less 98.8% 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 different fitted curves shown in Fig. 5. Hence, the fitted regression equation i.e., the predictive statistical model applicable for this data set is a cubic curve for reference examiner (Ex-46) given in the following equation and respective model diagram in Fig. 6, and with the help of this predictive model equation all the raw scores for all the examiners can be converted to their corresponding scaled scores.

$$\text{Raw score } (AR) = 15.356 + 7.18 \times 10^{-1}\text{Percentile} - 1.10 \times 10^{-2}\text{Percentile}^2 + 6.81 \times 10^{-5}\text{Percentile}^3$$

After getting all the Scaled Scores, it will be again transformed by the method of Origin Change (suitable, here it is 4.744) without hampering the relative position of the examinees. This is required towards legal aspects. This may be noted that the data corresponding to Nepali (medium of writing) (evaluated by EX61) is very less which does not conform to the assumption of Normality. Similarly all the 1238 examinees corresponding to all the eight examiners are being calculated. The comparison towards Raw scores and Scaled scores (thus obtained) are being represented by the Box plots shown in Fig. 7 & 8 and in following Table 5 to obtained the final merit list.

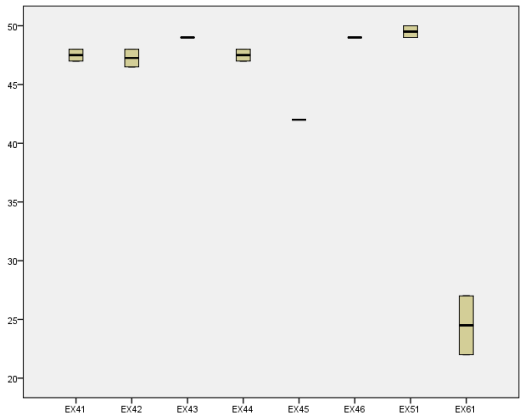


Fig. 7. Box-plots of Raw scores given by eight examiners.

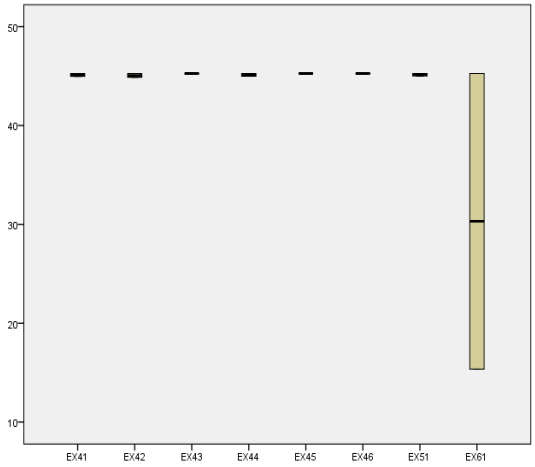


Fig. 8. Box-plots of Scaled scores given by eight examiners.

Finally one can shift the origin as required for the purpose of legal aspects without hampering the relative position of a candidate.

V. CONCLUSION

For a single subject paper of a descriptive type, which is judged by several examiners, the equi-percentile method can be used for removing the Examiners' Bias. This is also applicable to several test papers (mainly objective in nature) of different difficulty levels. But, in case of different test papers carrying different maximum marks required 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 be able to remove the examiners' bias. In all such cases, to solve the problem, the underlying distribution of marks awarded by different examiners is transferred to the distribution of the reference examiner through the process of converting raw scores to percentile scores could be adopted towards removing examiners' bias. The beauty is that these scaled scores are additive in nature, which enables one to prepare the final merit list.

Table 5: Comparison of distribution based conversion of raw score to scaled score for final merit list.

Can. ID	Marks Obtained	Individual Rank	Percentile	Scaled Scores	Examiner	Can. ID	Marks Obtained	Individual Rank	Percentile	Scaled Scores	Examiner
CANEX2 0010	50	1	100	45.256	EX51	CANEX2 0265	47	3	98.63945578	44.51001888	EX46
CANEX2 0068	49	1	100	45.256	EX43	CANEX2 1040	47	3	98.63945578	44.51001888	EX46
CANEX2 0313	49	1	100	45.256	EX43	CANEX2 0089	45	3	98.63013699	44.50502742	EX42
CANEX2 0239	49	1	100	45.256	EX46	CANEX2 0121	45	4	98.48484848	44.42741143	EX44
CANEX2 1022	49	1	100	45.256	EX46	CANEX2 0958	45	4	98.48484848	44.42741143	EX44
CANEX2 0050	48	1	100	45.256	EX41	CANEX2 0170	45.5	4	98.46153846	44.41499463	EX41
CANEX2 0599	48	1	100	45.256	EX42	CANEX2 0794	45.5	4	98.46153846	44.41499463	EX41
CANEX2 0131	48	1	100	45.256	EX44	CANEX2 0090	44.5	4	97.94520548	44.14248462	EX42
CANEX2 0081	42	1	100	45.256	EX45	CANEX2 0087	40	7	97.56097561	43.94281679	EX45
CANEX2 1116	42	1	100	45.256	EX45	CANEX2 0494	40	7	97.56097561	43.94281679	EX45
CANEX2 0933	27	1	100	45.256	EX61	CANEX2 0863	40	7	97.56097561	43.94281679	EX45
CANEX2 0127	49	2	99.56331878	45.01281437	EX51	CANEX2 1095	40	7	97.56097561	43.94281679	EX45
CANEX2 0247	49	2	99.56331878	45.01281437	EX51	CANEX2 1117	40	7	97.56097561	43.94281679	EX45
CANEX2 0310	49	2	99.56331878	45.01281437	EX51	CANEX2 1119	40	7	97.56098	43.94282	EX45
CANEX2 0312	49	2	99.56331878	45.01281437	EX51	CANEX2 0698	44	6	97.47475	43.89837	EX44
CANEX2 0324	49	2	99.56331878	45.01281437	EX51	CANEX2 1058	45	6	97.4359	43.87839	EX41
CANEX2 0812	49	2	99.56331878	45.01281437	EX51	CANEX2 0266	46	5	97.27891	43.79792	EX46
CANEX2 0926	47	2	99.49494949	44.97506326	EX44	CANEX2 1008	46	5	97.27891	43.79792	EX46
CANEX2 1161	47	2	99.49494949	44.97506326	EX44	CANEX2 0091	42	5	97.26027	43.7884	EX42
CANEX2 0177	47	2	99.48717949	44.97077846	EX41	CANEX2 0598	42	5	97.26027	43.7884	EX42
CANEX2 0108	46.5	2	99.31506849	44.87615545	EX42	CANEX2 1188	43	7	96.9697	43.6407	EX44
CANEX2 0031	41.5	3	99.18699187	44.80609891	EX45	CANEX2 0002	48	8	96.94323	43.62732	EX51
CANEX2 0060	46.5	3	98.97435897	44.6904617	EX41	CANEX2 0011	48	8	96.94323	43.62732	EX51
CANEX2 0511	41	4	98.7804878	44.58575454	EX45	CANEX2 0165	48	8	96.94323	43.62732	EX51
CANEX2 0800	41	4	98.7804878	44.58575454	EX45	CANEX2 0315	48	8	96.94323	43.62732	EX51
CANEX2 1128	41	4	98.7804878	44.58575454	EX45	CANEX2 0071	48	3	96.92308	43.61714	EX43

VI. FUTURE SCOPE

This methodology could be adopted to more of the examination data in connection with the recruitment process to judge the nature of conversion from raw scores to Scaled scores. This is of immense help to different administrators to remove the Examiner's Bias / Rating Bias, when there is more than one subject or sets of question paper for a recruitment process.

ACKNOWLEDGEMENTS

We are acknowledging the West Bengal Police Recruitment Board for constant support to use the required data sets for analysis. We also acknowledge the West Bengal University of Animal and Fishery Sciences for providing the computation and other facilities.

Conflict of Interest. The authors declare that they have no conflicts of interest.

REFERENCES

[1]. http://www.examscomp.com/wp-content/uploads/2014/12/Equalisation-of-scores-in-SSC-Examinations.pdf

[2]. Lawton, M., Hu, M. T., Baig, F., Ruffmann, C., Barron, E., Swallow, D. M., & Williams, N. (2016). Equating scores of the University of Pennsylvania Smell Identification Test and Sniffin' Sticks test in patients with Parkinson's disease. *Parkinsonism & related disorders, 33*, 96-101.

[3]. Lee, E. (2013). Equating multidimensional tests under a random groups design: A comparison of various equating procedures.

[4]. 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.

[5]. Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. Springer Science & Business Media.

[6]. 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.

[7]. Holland, P. W., & Dorans, N. J. (2006). Linking and equating. *Educational measurement, 4*, 187-220.

[8]. Sahu, S &Sahu, S, (2020). Prof. Sahu's Methodology of Distribution Dependent Equalization of Scores to Remove Examiner's Bias and / or Difficulty Bias, (Registration No: L-89634/2020 Dated: 17/02/2020, Serial No. 714, e-register: February 2020, copyright.gov.in)

How to cite this article: Sahu, S., Harirajan, K., Mahapatra, G. and Sahu, S. (2020). Equalization of Scores through Statistical Modeling towards Removing Examiners' Bias for Arithmetic and Reasoning Paper. *International Journal on Emerging Technologies*, *11*(2): 295–303.