Journal of Research and Reflections in Education
December 2016, Vol. 10, No.2, pp 156- 168
http://www.ue.edu.pk/jrre

High-Stake Testing in Punjab: Inter-rater Reliability in the Scoring of Secondary School Certificate (SSC) Examination

[1] Sehar Rashid, [2] Nasir Mahmood

[1] PhD (Scholar) Institute of Education and Research Quaid-i-Azam Campus, University of the Punjab Lahore,
[2] Professor and Dean Faculty of Education, Allama Iqbal Open University

(Email: sehar_rashid12@yahoo.com)

The study aimed to assess inter-rater consistency in the scoring of sub-examiners of SSC. The study aimed to examine the level of inter-rater reliability as a source of variation in measurement in the scoring of SSC papers of high-stake testing. The population comprised sub-examiners of BISE Lahore for subject of English and Urdu for session 2013-2014. Stratified random sampling technique was used which identified 42 strata in population. The number of sub-examiners from 26 representatively selected strata was 98. The instrument of the study was four solved papers of annual examination for each subject. For data analysis, consistency estimates of inter-rater reliability approach were used that include Spearman correlation. The results were benchmarked according to Landis and Koch-Kappa's benchmark scale. The results revealed moderate inter-rater reliability in scoring of the secondary examination of high-stake testing. There is an evident need on the part of the examination boards to inbuilt measures to improve inter rater reliability in scoring to improve trust in measurement and considering far reaching implication in subsequent use of high stake examinations.

Key words: *inter-rater reliability, high-stake testing, consistency estimates of inter-rater reliability, Secondary School Examination*

Introduction

Since independence of Pakistan, University of the Punjab was responsible to conduct examination at secondary and higher-secondary level in Pakistan. First examination board named 'Board of secondary education, Punjab' was established in 1954 in the province of Punjab. Currently, nine boards are working in different divisions of Punjab aimed "to conduct examination in fair, unbiased, transparent and judicious environment so that our educated nationals may be able to contribute meaningfully in the competitive world" (BISE, 2016).

The education system in Pakistan consists of primary, elementary, secondary, higher secondary and tertiary level. To award the certificate of secondary level, an examination called Secondary School Certificate (SSC) examination is conducted by different boards of intermediate and secondary education (BISEs). The significance of SSC examination cannot be neglected as the future of students' employability or higher education lie on the results of these examinations.

Large numbers of candidate get registered for appearing in SSC examination held usually in March-April every year in the province of Punjab. The results of SSC examination is announced in usually July-August every year. Thus, there is a gap of four months between the administration and result announcement. This time period consumed for the collection of papers to the paper checking centers, coding of the papers for scoring purpose, allotment of papers to the sub-examiners for scoring, scoring of the papers by sub-examiners, rechecking by the superintendent officer, compilation of the results and printing of the results. The calculated time for the scoring of the papers is only about two months. Thus, it is not possible for a single rater to score all

papers for a subject in such a short period of time. As a result, it is mandatory to involve multiple raters for the scoring of papers for each subject. Whenever multiple raters take part in scoring of papers, the subjectivity is involved in a number of ways as human being are notorious for their inconsistency, can easily distracted, get tired from repetitive tasks, can misinterpret or day dream (Trochim, 2006).All these factors influenced the reliability and consistency of the results. This creates inter-rater reliability threat leading to suspect the credibility of the results produced by sub-examiners.

To overcome the threats to the accuracy of measurement, assessment procedure followed by examination boards must be according to some standard or policy to have sound tests, scored properly and used appropriately. This concern has been debated globally as some of the prominent measurement organizations such as American Education Research Association (AERA), the American Psychological Association (APA) and the National Council on Measurement and Evaluation (NCME) have all individually developed policy or position statement about high-stake testing that are remarkably similar (Kubiszyn & Borich, 2003). By reviewing the educational policies of Pakistan (Govt. of Pakistan, 1947, 1970, 1972, 1998, 2009-10), it can be understood that educational assessment has remained a subject of interest since its inception. Accordingly it can be assumed that importance of high-stake testing has been approved globally. Now, the debate is 'what is the best way to be adopted by high-stake testing'. In Pakistan, trust in measured student achievement has been seen with suspect by parents, teachers and educational institutions. Such as after the announcement of results, number of applications received by the examination boards for paper rechecking every year which give an impression of dissatisfaction (Haider, July 30, 2013). This might be because of the reason that candidates got low scores as compared to the expectations or optimum effort by those candidates. Noticeably the number of application for rechecking of the papers increased proportionately every year. The cases of those scoring below their expectation are pointed out in case they file a request for re-checking but those who receive scores higher than expectations are never pointed out.

The inter-rater reliability is the level to which a student obtains the same scores if different raters scored the performance (Nitko, 1996). The measure of inter-rater reliability has been researched globally since last century. There are few researches made nationally to study the examination system of the country (Shah, 1998; Bashir, 2002; Shirazi, 2004; Kiani, 2004; Jaffri, 2006; Jilani, 2009). The focus of these researches was to assess the quality of whole examination system but not scoring reliability, specifically inter-rater reliability of scoring of the sub-examiners. So the study was aimed to initiate the concern to assess inter-rater reliability of sub-examiners in SSC examination. However, the study was delimited to the subjects of Urdu and English as these subjects have essay type questions which can be best check for examining inter-rater reliability, and both subjects hold a prominent place in the study of all students as compulsory subjects.

This study was aimed to examine level of inter-rater reliability as a source of measurement error in scoring of SSC papers high-stake testing. The objectives were attained by dealing with the questions: what is the level of inter-rater reliability in the scoring of the sub-examiners for SSC papers of BISE Lahore? and, how much inter-rater reliability affects the measurement in scoring of SSC papers of BISE Lahore on the base of gender, qualification and cadre of the sub-examiners?

Literature Review

The reliability of scoring is significant aspect to ensure quality control of assessment procedure that affects candidate's life chance. A major characteristic of reliability of scoring is the inter-rater reliability. It is useful to endorse the fairness of the criteria of scoring and to uphold clear understanding among raters. Inter-rater reliability has been researched for all grades of education across several subjects and assessment methods. Meadows and Billington (2005) noticed that earliest studies focused to determine the inter-rater reliability in scoring of high school teachers. The earliest study by Starch and Elliot (1912) found problems related with the reliable assessment of English. In another study, Starch and Elliot (1913a) found low levels of inter-rater reliability of test scripts of geometry and history.

Several research studies have reported the inter-rater reliability for different subjects at higher education level. Primarily, Hartog and Rhodes (1936) assessed the range of agreement among couples of raters for the subject of history which is from -0.41 to 0.85 with an average of just 0.44. This indicates a low inter-rater reliability among the raters. Murphy (1978, 1982) conducted a series of studies for in-depth analysis of scoring reliability for the subject of English (level A) and discussed that reliability in scoring of rater will increase by increasing the quantity of components. Recently, a research by Porterand and Jelinek (2011) found the range of inter-rater reliability from poor to moderate.

It has been debated before 1970s that inter-rater reliability may be irrespective of the subject area being scored. But, in 1970s it was understood that inter-rater reliability is reliant on the subject area being evaluated (Byrne, 1979). James (1974) and McVey (1975) found that the inter-rater reliability was high in subjects like physics and electronic engineering

It is generally considered that subjectivity of the raters highly affects the essay type material. Therefore, measure of inter-rater reliability can be better assessed for the essay type material. Finlayson (1951) found the reliability coefficient for the team of scorer and claimed that reliability of essay is well assessed by applying test re-test reliability. It infers that measure of inter-rater reliability for essays also affected by the measurement technique being used to assess it. Bryne (1979) concluded that essay questions offered the extreme reliability problem irrespective of the subject area.

Approaches to measure inter-rater reliability of scoring in high-stake testing. Generally inter-rater reliability is the uniformity among two or more examiners assesses the identical documents by the use of same scoring scheme (Bailey 1998) at a specific time (Stemler 2004). To deal with the question of measurement of the inter-rater reliability, literature provides three approaches i.e. (a) inter-rater agreement also known as consensus estimates of reliability; (b) measure of inter-rater reliability and (c) consistency estimates of inter-rater reliability. All of these three have different assumptions with its advantages and disadvantages and different measurement techniques. A brief introduction to all of these techniques is described here.

Inter-rater agreement. This approach is known as consensus estimate of reliability with the supposition that eligible sub-examiners should be competent to make an exact agreement regarding the application of different levels of a scoring scheme for the same phenomenon. If two sub-examiners make a similar agreement about the use of scoring scheme, they are supposed to share a collective explanation of the construct (Stemler, 2004). The raters essentially have to be trained to point of exact agreement.

This approach is said to be appropriate for nominal data and the diverse stages on the scoring scheme denote descriptively diverse thoughts and linear continuum.

The most general statistical technique to measure an agreement approximation is made by taking the percentage of agreement. This technique is instinctively interesting, easy to compute, and easy to describe but it is likely to attain greater percent-agreements as maximum values collapse in one sort of the scoring scheme (Hayes & Hatch, 1999). In contrast, this technique is time taking and labor intensive.

Another consensus estimate is Cohen's (1960) kappa coefficient. This statistics considers the degree of agreement expected by chance and amends the percentage of estimated agreement accordingly. A benchmark for kappa coefficient proposed by Landis and Koch (1977) The foremost benefit of this statistic is for those who concern about the falsely inflated value of percentage of agreement. Its main drawback is that it is tough to interpret. This technique is applied only for two raters.

According to Stemler (2004), this approach is advantageous as its calculations can be easily done manually and can be useful in identifying problems with raters' understandings about the use of scoring scheme. A high value of agreement infers that sub-examiners are fundamentally producing the same results. On other hand, this approach is disadvantageous because statistics to measure inter-rater reliability as it must be calculated independently for each item and for each pair of raters. An extensive volume of energy and time is

needed to train the sub-examiners so that they make consensuses. Training of sub-examiners to make a forced agreement lessens the impartiality of scoring which is a threat to validity of results produced (Linacre, 2002).

Measure of inter-rater reliability. This technique requires information by all sub-examiners about summary scores independently for each sub-examiner. Linacre (2002) has argued "*it is the accumulation of information, not the rating themselves, that is decisive"* (p. 858). Accordingly, each sub-examiner is supposed to provide some distinctive information which will beneficial in producing a summary score independently for each sub-examiner. So it is not essential for two sub-examiners to agree about the application of scoring scheme.

This approach is most beneficent when diverse rank of scoring scheme are proposed to characterize varied ranks of basic one-dimensional concept. Under this approach, inter-rater reliability can be measured by different statistical techniques. Common techniques are factor analysis (Harman, 1967), generalizability theory(Shavelson& Webb, 1991) and many-facets Rasch model (Linacre,1994).

According to Stemler (2004), this approach is used for ordinal data and is advantageous in the way that the measures can be taken as interpretation inaccuracies for each sub-examiner or a group of sub-examiners. It can efficiently hold scoring of numerous sub-examiners by simultaneously calculating measures across all scored items. Sub-examiners are not needed to score entire items in order to arrive at a measure of inter-rater reliability.

Consistency estimates of inter-rater reliability. This approach is based on the assumption that two sub-examiners are not needed to understand the scoring scheme similarly, given that each sub-examiner is reliable in categorizing the condition or occasion as per own understanding about scoring scheme. This approach is used for continuous data.

Under this approach, commonly used statistics to calculate the reliability among sub-examiners is Pearson correlation coefficient that can only be applied to the normally distributed data for two sub-examiners for one item at a time. Another statistic is Spearman's rank coefficient that can be applied for normally distributed data when all raters provide rating for all cases. One more statistical technique is Cronbach's alpha coefficient which is used to understand the degree to which the scoring of two sub-examiners measures a mutual dimension.

Stemler (2004) proposed that this approach is advantageous in the way that raters are not needed to be trained to have a similar agreement with each other. Statistics applied under this approach permits for general estimate of consistency between numerous sub-examiners. It is disadvantageous in the way that sub-examiners may vary gradually in scoring they applied and vary in categorization of scoring scheme they used. This measure is extremely delicate to the distribution of the data.

It is difficult to be conclusive about mostly used statistics for inter-rater reliability as Hallgren (2012) reported that most of the studies failed to report the most appropriate statistics for inter-rater reliability. He argued that most of the studies apply percentage agreement regardless of the objectives of the study because of its ease to compute. These are the weaknesses of most of the studies in literature which assess inter-rater reliability. However, not only one method is the answer of all issues. Thus, the selection of the statistics for inter-rater reliability should rely on the objectives of the study, nature of the data and sub-examiners participating in the study.

Methodology

The research design for this study was descriptive. A cross-sectional survey was conducted for data collection from a group of people to describe the aspects and characteristics of population.

Population. The population comprised 539 sub-examiners of Urdu subject and 345 sub-examiners of English subject who were involved in SSC paper checking for BISE, Lahore annual examination 2014. The lists of those sub-examiners were collected from the BISE, Lahore.

Sample and sampling technique. The lists of sub-examiners collected from BISE Lahore illustrate different characteristics of the sub-examiners, so to have representation of all the characteristics of the population stratified random sampling technique was used for the study. The

sampling technique was applied in three stages. At first stage, sub-examiners were stratified on the basis of their gender (male and female). At second stage, sub-examiners of each gender were further stratified on the basis of their qualification categorized as graduation (B.A/ B.Sc) and post-graduation (M.A/ M.Sc). At third stage, sub-examiners of each strata based on gender and qualification were further stratified on the basis of their cadres which were elementary school teacher (EST), secondary school teacher(SST), Arabic teacher (A.T), retired teacher (Rtd.), Teacher of registered private school (T).

After applying this technique, there was 42 (22 for Urdu and 20 for English) strata of sub-examiners for the selected subjects. Sample of 98 sub-examiners participated in the study willingly. Distribution of sample according to their corresponding stratum is given in table 1.

Table 1

Distribution of sample by gender, qualification and cadre

	Category of sub-examiners	No. of raters for Urdu	No. of raters for English
Male	MA/MSc, EST	07	06
	MA/MSc, SST	05	04
	MA/MSc, T	03	03
	MA/MSc, A.T	03	----
	MA/MSc, Rtd.	04	----
	BA/BSc, EST	10	03
	BA/BSc, SST	04	03
	BA/BSc, T	04	03
	BA/BSc, A.T	03	----
	BA/BSc, Rtd.	03	03
Female	MA/MSc, EST	03	03
	MA/MSc, SST	03	03
	MA/MSc, T	03	03
	BA/BSc, SST	03	03
	BA/BSc, Rtd.	----	03
	Total no. of raters (98)	58	40

BA Bachelor of Arts, *BSc* Bachelor of Science, *MA* Master of Arts, *MSc* Master of Science, *EST* Elementary School Teacher, *SST* Secondary School Teacher, *A.T* Arabic Teacher, *Rtd.* Retired Teacher, *T* Teacher of registered private school

Instrument

Solved scripts of Urdu and English 2014.The actual scripts were not provided by the BISE Lahore due to the secrecy reasons. Thus, the same papers (for both subjects) were administered within two weeks after the annual secondary school certificate examinations of BISE Lahore, session 2013-2014 to four students for each subject. These students had already appeared in the annual examination. The solved papers for both subjects were comprised of restricted response items and extended response items of the question papers. The part of supply type items of the question paper for both subjects were not administered to the students as the scoring subjectivity is irrelevant for such items All those eight scripts were used as an instrument for the sub-examiners of English and Urdu respectively. Thus we can say that scripts were test-retest papers but the influence of test-retest is not related to determination of inter-rater reliability as the study concerned with the scoring of sub-examiners on those scripts instead of performance of students on those scripts.

Data collection. Owing to the manageable size of the sample and reducing the effect of sub-examiners, we collected data from the subjects of the study. For the purpose of data collection, we traced the contact numbers of the sub-examiners mentioned in the lists provided by BISE Lahore and contacted them to take their consent to participate in this study. After getting consents from the sub-examiners who were willing to participate in the study, we visited to

the sub-examiners as per the appointment taken after contacting them. Each of the respondents was requested to provide their scoring for four papers (Urdu or English) of the subject for which he/she score BISE examinations.

Data analysis. A consistency estimate of inter-rater reliability was selected approach to measure inter-rater reliability in scoring of sub-examiners for this research. According to this approach, there are three statistics which can be applied to measure inter-rater reliability in the scoring i.e. Cronbach alpha applied, Pearson's correlation coefficient and Spearman's correlation coefficient.

Gwet (2012) stated that *Spearman's correlation is more appropriate for evaluating inter-rater reliability than Pearson's correlation due to the fact that if the 2 series of ratings are in agreement with respect of ranking of subjects, then their relationship needs not be linear for Spearman's correlation to be high (p. 251).*

Thus the Spearman's correlation coefficient was selected to measure inter-rater reliability in scoring of the sub-examiners for this study. Some of the non-parametric statistical techniques such as Mann-Whitney U test and Kruskal-Wallis test were also applied for the comparison of consistency in scoring of group of sub-examiners.

Benchmark scale for interpretation of variance in consistency. Benchmarking is important to interpret the magnitude of the correlation coefficient for inter-rater reliability. There are various benchmarks scales proposed in the literature. The most precise benchmark scale proposed by the Landis and Koch (1977) which was selected to interpret the magnitude of the correlation coefficient for inter-rater reliability in scoring of sub-examiners (Table 2).

Table 2

Landis and Koch-Kappa's benchmark scale

Kappa Statistics	Strength of Agreement
< 0.0	Poor
0.0 to 0.20	Slight
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Substantial
0.81 to 1.00	Almost perfect

Results and Conclusions

The inter-rater reliability in the scoring of the sub-examiners is described as the range of variance in inter-rater reliability for a number of comparisons for each category. Murphy (1982) argues that the simplest approach of labeling the extent of variation because of different examiners doing scoring is the average value. Thus the mean value is also computed for the variance in the scoring of the examiner for each category of sub-examiners.

Measure of inter-rater reliability in scoring of sub-examiners Inter-rater reliability in scoring of sub-examiners was measured by applying Spearman rho for the scoring of sub-examiners.

Table 3

Inter-rater reliability among male sub-examiners for the subject of Urdu

	Category of sub-examiners	No. of raters	Range of variance in consistency (r)	Mean
Male	MA/MSc, EST	7	0.451-0.834	0.63
	MA/MSc, SST	5	0.382-0.745	0.55
	MA/MSc, T	3	0.439-0.457	0.448
	MA/MSc, A.T	3	0.322-0.468	0.39
	MA/MSc, Rtd	4	0.502-0.763	0.67
	BA/BSc, EST	10	0.423-0.970	0.64
	BA/BSc, SST	4	0.438-0.687	0.52
	BA/BSc, A.T	3	0.462-0.517	0.51
	BA/BSc, T	4	0.558-0.715	0.65
	BA/BSc, Rtd.	3	0.667-0.738	0.71

**Correlation is significant at 0.01 level (2-tail)

Table 3reveals that agreement for scoring among male raters with qualification MA/MSc and cadre 'A.T'; and, male raters with qualification MA/MSc and cadre 'T' not varied extensively. However, the mean value of the consistency were 0.39 and 0.44 among the male raters with qualification of MA/MSc and cadre 'A.T'; and, male raters with qualification of MA/MSc and cadre 'T' respectively, which shows fair inter-rater reliability in scoring of papers of Urdu.

Table 3also reveals that agreement for scoring among male raters with qualification BA/BSc and cadre 'SST' varied in a small range with difference of 0.11; scoring among male raters with qualification MA/MSc and cadre 'SST' varied in a wide range with difference of 0.36; scoring among male raters with qualification BA/BSc and cadre 'A.T' varied in a small range with difference of 0.10; and, scoring among male raters with qualification BA/BSc and cadre 'T' varied in a medium range with difference of 0.24. However, the mean values of the consistency were 0.52, 0.55, 0.51 and 0.65 among the male raters with qualification of BA/BSc and cadre 'SST'; male raters with qualification of MA/MSc and cadre 'SST'; male raters with qualification of BA/BSc and cadre 'A.T'; and, male raters with qualification of BA/BSc and cadre 'T' respectively, which shows moderate inter-rater reliability in scoring of papers of Urdu.

Table 3 shows that agreement for scoring among male raters with qualification MA/MSc and cadre 'Rtd.' varied in medium range with difference of 0.26; scoring among male raters with qualification MA/MSc and cadre 'EST' varied in a wide range with difference of 0.38; scoring among male raters with qualification BA/BSc and cadre 'EST' varied in a wide range with difference of 0.55; and, scoring among male raters with qualification BA/BSc and cadre 'Rtd.' not varied extensively. However, the mean values of the consistency are 0.67, 0.63, 0.64 and 0.71 among the male raters with qualification of MA/MSc and cadre 'Rtd.'; male raters with qualification of MA/MSc and cadre 'EST' and, male raters with qualification of BA/BSc and cadre 'EST; and, male raters with qualification of BA/BSc and cadre 'Rtd.' respectively; which shows substantial inter-rater reliability in scoring of papers of Urdu.

Table 4

Inter-rater reliability among female sub-examiners for the subject of Urdu

Category of sub-examiners		No. of Raters	Range of variance in consistency (r)	Mean
Female	MA/MSc, EST	3	0.383-0.750	0.51
	MA/MSc, SST	3	0.401-0.494	0.45
	MA/MSc, T	3	0.697-0.794	0.75
	BA/BSc, SST	3	0.506-0.594	0.56

**Correlation is significant at 0.01 level (2-tail)

Table 4 reveals that agreement for scoring among females with qualification MA/MSc and cadre 'EST' varied in a wide range with difference of 0.37; scoring among female raters with qualification MA/MSc and cadre 'SST'; and, female raters with qualification BA/BSc and cadre 'SST' not varied extensively. However, the mean values of the consistency were 0.51, 0.45 and 0.56 among the female raters with qualification of MA/MSc and cadre 'EST'; female raters with qualification of MA/MSc and cadre 'SST'; and, female raters with qualification of BA/BSc and cadre 'SST' respectively, which shows moderate inter-rater reliability in scoring of papers of Urdu.

Table 4also reveals that agreement for scoring among female raters with qualification MA/MSc and cadre 'T' varied in a small range with difference of 0.10. However, the mean value of the consistency was 0.75 among the female raters with qualification of MA/MSc and cadre 'T', which shows substantial inter-rater reliability in scoring of papers of Urdu.

Table 5

Inter-rater reliability among male sub-examiners for the subject of English

Category of sub-examiners		No. of Raters	Range of variance in consistency (r)	Mean
Male	MA/MSc, EST	6	0.548-0.991	0.82
	MA/MSc, SST	4	0.849-0.926	0.89
	MA/MSc, T	3	0.610-0.642	0.62
	BA/BSc, EST	3	0.532-0.758	0.65
	BA/BSc, SST	3	0.599-0.840	0.71
	BA/BSc, T	3	0.668-0.700	0.68
	BA/BSc, Rtd	3	0.850-0.879	0.86

**Correlation is significant at 0.01 level (2-tail)

Table 5 shows that agreement for scoring among male raters with qualification MA/MSc and cadre 'EST' varied in wide range with difference of 0.45; scoring among male raters with qualification MA/MSc and cadre 'SST'; and, male raters with qualification BA/BSc and cadre 'Rtd.' not varied extensively. However, the mean values of the consistency were 0.82, 0.89 and 0.86 among the male raters with qualification of MA/MSc and cadre 'EST'; male raters with qualification of MA/MSc and cadre 'SST'; and, male raters with qualification of BA/BSc and cadre as a 'Rtd.' respectively, which shows almost perfect inter-rater reliability in scoring of papers of English.

Table 5 also reveals that agreement for scoring among male raters with qualification BA/BSc and cadre 'EST'; and, male raters with qualification BA/BSc and cadre 'SST' varied in

medium range with difference of 0.25 and, scoring among male raters with qualification BA/BSc and cadre 'T' not varied extensively. However, the mean values of the consistency are 0.65, 0.71 and 0.68 among the male raters with qualification of BA/BSc and cadre 'EST'; male raters with qualification of BA/BSc and cadre 'SST'; and, males raters with qualification of BA/BSc and cadre 'T' respectively, which shows substantial inter-rater reliability in scoring of papers of Urdu.

Table 5 also reveals that agreement for scoring among male raters with qualification MA/MSc and cadre 'T' not varied extensively. However, mean value of the consistency among raters was 0.62 between the male raters with qualification of MA/MSc and cadre 'T', which shows moderate inter-rater reliability in scoring of papers of English.

Table 6

Inter-rater reliability among male sub-examiners for the subject of English

Category of sub-examiners		No. of Raters	Range of variance in consistency (r)	Mean
Female	MA/MSc, EST	3	0.589-0.751	0.68
	MA/MSc, SST	3	0.627-0.760	0.71
	MA/MSc, T	3	0.672-0.816	0.72
	BA/BSc, SST	3	0.598-0.714	0.66
	BA/BSc, Rtd.	3	0.724-0.867	0.80

**Correlation is significant at 0.01 level (2-tail)

Table 6 reveals that agreement for scoring among female raters with qualification BA/BSc and cadre 'Rtd.' varied in a small range with difference of 0.14. However, the mean value of the consistency was 0.80 among the female raters with qualification of BA/BSc and cadre 'Rtd.', which shows almost perfect inter-rater reliability in scoring of papers of English.

Table 6 also reveals that agreement for scoring among female raters with qualification MA/MSc and cadre 'EST' varied in a small range with difference of 0.17; scoring among female raters with qualification MA/MSc and cadre 'SST', and, female raters with qualification MA/MSc and cadre 'T' varied in a small range with difference of 0.14; scoring among female raters with qualification BA/BSc and cadre 'SST' varied in a small range with difference of 0.13. However, the mean values of the consistency were 0.68, 0.71, 0.72 and 0.66 among the female raters with qualification of MA/MSc and cadre 'EST'; female raters with qualification of MA/MSc and cadre 'SST'; female raters with qualification of MA/MSc and cadre 'T', and, female raters with qualification of BA/BSc and cadre 'SST' respectively, which shows substantial inter-rater reliability in scoring of papers of English.

Measure of inter-rater reliability of scoring on the basis of gender. Inter-rater reliability in scoring of sub-examiners on the basis of their gender was measured by applying Mann-Whitney U test on mean *rho* value of sub-examiners.

Table 7

Subject	Gender	N	Mean Rank	U-value	Z	sig
Urdu	Male	46	32.8	124.0	-2.918	0.004
	Female	12	16.8			
English	Male	25	23.8	103.0	-2.361	0.018
	Female	15	14.8			

Comparison among sub-examiners scoring on the basis of gender N=98

Table 7 shows that there is a significant difference in inter-rater reliability value across male and female sub-examiners for the subjects of Urdu and English. The mean rank value revealed that male sub-examiners are more consistent in their scoring as compare to the female sub-examiners for the subjects of Urdu and English.

Measure of inter-rater reliability of scoring on the basis of Qualification. Inter-rater reliability in scoring of sub-examiners on the basis of their qualification (graduation/ post-graduation) was measured by applying Kruskal-Wallis test on mean *rho* value of sub-examiners.

Table 8

Comparison among sub-examiners' scoring on the basis of qualification

Subject	Qualification	N	Mean Rank	U-value	Z	Sig
Urdu	MA/MSc	31	21.6	176.0	-3.78	0.000
	BA/BSc	27	38.4			
English	MA/MSc	22	21.6	173.0	-.680	0.497
	BA/BSc	18	19.1			

N=98

Table 8 shows that there is no significant difference in inter-rater reliability value across graduate qualified sub-examiners and post-graduate qualified sub-examiners for the subject of English but there is a significant difference in inter-rater reliability value across graduate qualified sub-examiners and post-graduate qualified sub-examiners for the subject of Urdu. The mean rank value revealed that graduate sub-examiners are more consistent in their scoring as compare to the postgraduate sub-examiners for the subject of Urdu.

Measure of inter-rater reliability of scoring on the basis of cadre. Inter-rater reliability in scoring of sub-examiners on the basis of their cadres was measured by applying Kruskal-Wallistest on mean *rho* value of sub-examiners.

Table 9

Comparison among sub-examiners' scoring on the basis of cadre

Subject	Cadre	N	Mean Rank	Chi-Square	df	Sig
Urdu	Elementary School Teacher	20	39.8	30.248	4	.000
	Secondary School Teacher	15	16.9			
	Teacher of Registered school	10	28.5			
	Arabic Teacher	6	9.6			
	Retired teacher	7	45.29			
English	Elementary School Teacher	12	20.17	13.698	3	.003
	Secondary School Teacher	13	19.57			
	Teacher of Registered school	9	12.56			
	Retired teacher	6	35.17			

N=98

Table 9 shows that there is a significant difference in inter-rater reliability value across sub-examiners with cadre elementary school teachers, secondary school teachers, teachers of registered schools and retired teachers for the subject of Urdu and English. The mean rank value revealed that sub-examiners with cadre of retired teachers are highly consistent in their scoring and sub-examiners with cadre of Arabic teachers are least consistent in their scoring for the subject of Urdu. The mean rank value revealed that sub-examiners with cadre of retired teachers are highly consistent in their scoring and sub-examiners with cadre of 'teachers of registered school' are least consistent in their scoring for the subject of English.

Discussion

The findings of the study revealed that a same paper can get 30 to 90 scores for Urdu subject and same paper can get 50 to 90 scores for English subject if scored by different sub-examiners. This shows that the results of a candidate merely depend upon who is scoring his/her paper. It is alarming for the credibility of the results produced by high-stake testing. These lead us to a major conclusion of this study that there is a moderate inter-rater reliability in scoring of SSC papers of BISE, Lahore. This is a major threat to the credibility of the results produced by the high stake testing. A big mound of researches (Starch & Elliot,1912; Hartog & Rhodes, 1936; Finlayson, 1951; Murphy, 1978, 1982; James, 1974; McVey, 1975; Byrne, 1979) also identified low inter-rater reliability of the high-stake testing which support the finding of this research. This threat to the credibility of the results of high-stake testing in Punjab might be reduced by monitoring the inter-rater reliability of the sub-examiners individually. It might be done in a training session before the start of scoring session by opting the scoring of sub-examiners on the same sample paper and then measuring inter-rater reliability in the scoring of sub-examiners. It is recommended to allowing scoring only to those sub-examiners who possess high inter-rater reliability in their scoring at the end of training session.

The findings of the study revealed that the men sub-examiners were more consistent in their scoring as compared to the women sub-examiners for the subjects of Urdu and English. Greatorex and Bell (2002) support this findingthat men are more consistent in their scoring as compare to women sub-examiners. It might reflect men's sincerity towards their job as they have responsibility of their falimies to support financially. On the other hand, females' involve themselves in such jobs just to keep them busy or pass time.Thus,male sub-examiners might be preferred over female sub-examiners.

Other finding revealed that there is no significant difference in inter-rater reliability value across graduate qualified sub-examiners and post-graduate qualified sub-examiners for the subject of English but graduate sub-examiners are more consistent in their scoring as compared to the post graduate sub-examiners for the subject of Urdu. This finding concluded that a criterion of minimum qualification for the scoring of papers is enough to get high inter-rater reliability in the scoring of sub-examiners.

It was also found that sub-examiners with cadre of retired teachers are highly consistent in their scoring for the both subjects as compared to other cadres. This might be because the retired teachers have more time to spend on scoring as they do not have any other job and they have more relevant experience as compared to the in-service sub-examiners. So they score the papers willingly and aspiringly. While the other teachers have their job responsibilities and they undertake such tasks for the purpose to make money. Thus, retired teachers might be preferred for scoring.

Elementary school teachers were found second highly consistent in their scoring after retired teacher for both the subjects. This can be explained in terms that ESTs might feel privileged when they selected to score the scripts of secondary level. This might make them aspire to score scripts attentively as compared to SSTs. Thus it is recommended to involve more elementary school teachers who are eligible for scoring.

Findings of the present study score a question on all the measures executed to improve the credibility of the results by the BISE, Lahore. It is also a big threat to the life chance of all the candidates who appear such examinations. Thus there is a need to consider the measure of inter-rater

reliability by high-stake testing before the start of scoring session.

References

Bailey, K. M. (1998). *Learning about language assessment.* Pacific Grove : Heinle & Heinle Publishers.

Bashir, M. (2002) *A Study of Examination system of Pakistan and Development of a Model for Twenty First Century,* Ph.D. dissertation, Retrieved from *http://eprints.hec.gov.pk/6657/*

Board of Intermediate and Secondary Education Lahore, (2016) *Establishment and vision statement,*

Retrieved from *http://www.biselahore.com/about-us.htm*

Byrne, C. (1979) Tutor-marked assignments at the Open University: A question of reliability, *Teaching at a Distance, 15,* 34-43.

Cohen, J. (1960) A coefficient of agreement from nominal scale, *Educational Psuchological Meaurements, 20,* 37-46. doi:10.1177/001316446002000104

Finlayson, D. S. (1951) The reliability of the marking of essay, *British Journal of Educational Psychology, 21*(2), 126-134. doi:10.1111/j.2044-8279.1951.tb02776.x

Govt. of Pakistan.(1947) *The Pakistan Educational Conference*, Education Division, Karachi. p, 19.

Govt. of Pakistan.(1970) National Sub-Commission on Education Reforms, *Ministry of Education, Islamabad*. p. 49-51

Govt. of Pakistan.(1972) National Education Policy, *Ministry of Education, Islamabad*.

Govt. of Pakistan.(1998) National Education Policy, *Ministry of Education, Islamabad*.

Govt. of Pakistan.(2009-10) National Education Policy, *Ministry of Education, Islamabad*.

Greatorex, J. & Bell, J. F. (2002). *Does the gender of examiners influence their marking?* Paper presented at the Learning communities and assessment cultures: Connecting research with practice, University of Northumbria.

Gwet, K. L. (2012) *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters,* (3rd, Ed.) United States of America: Advanced Analytics, LLC.

Haider, A. (2013, July 30) .Marks rechecking mechanism displeases students, *Daily Times,* pp. 13, 6

Hallgren, K. A. (2012) Computing inter-rater reliability for observational data: An overview and tutorial, *Tutor Quant Methods Psychol, 8*(1), 23-34.

Harman, H. H. (1967) *Modern factor analysis,* Chicago: University of Chicago Press.

Hartog, P., & Rhodes, E. C. (1936) *The marks of examiners*, Landon: Macmillan and Co.

Hayes, J. R., & Hatch, J. A. (1999) Issues in measuring reliability, *Written Communication, 16*(3), 354-367. doi:10.1177/0741088399016003004

Jaffri, S. I. H. (2006) *A study of the effectiveness of boards examinations in Physics as reflected in the results of Boards of Intermediate and Secondary Education in Sindh,* P.hd dissertation, Retrieved from *http://eprints.hec.gov.pk/3403/*

James, C. (1974) The consistency of marking a physics examination , *Physics Education*, 9, 271-274.

Jilani, R. (2009) Problematizing high school certificate exam in Pakistan, A Washback perspective. *The Reading Matrics, 9*(2), 175-183.

Kiani, M., A. H. (2004) *A study to evaluate the examination system at Grade-V in the Punjab, based on Solo Taxonomy*, Doctoral dissertation, Retrieved from *http://prr.hec.gov.pk/Thesis/774S.pdf*

Kubiszyn, T., & Borich, G. (2003) *Educational testing and measurement, classroom applications and practice* (7th ed.), United States of America: John Wiley & Sons, Inc.

Landis, J. R., & Koch, G. (1977) The measurement of observer agreement for categorical data, *Biometrics, 33*(1), 159-174.

Linacre, J. M. (1994) *Many-facet Rasch measurement*,Chicago: MESA Press.

Linacre, J. M. (2002) Judge ratings with forced agreement, *Rasch Measurement Transactions, 16*(1), 857-858.

McVey, P. J. (1975) The error in marking examination script in electronic engineering, *International Journal of Electronic Engineering Education, 12*, 203-216.

Meadows, M., & Billington, L. (2005) *A review on the literature on marking reliability*, National Assessment Agency.

Murphy, R. J. (1978) Reliability of marking in eight GCE examnination, *British Journal of Educational Psychology*, 48(2), 196-200. doi:10.1111/j.2044-8279.1978.tb02385.x

Murphy, R. J. (1982) A further report of investigations into the reliability of marking of GCE examination, B*ritish Journal of Educational Psychology*, 52(1), 58-63. doi:10.1111/j.2044-8279.1982.tb02503.x

Nitko, A. J. (1996) *Educational assessment of students* (2nd ed.), United States of America:

Printice-Hall, Inc.

Porter, J. M. &Jelinek, D. (2011) Evaluating Inter-rater Reliability of a National Assessment Model for Teacher Performance, *International Journal of Educational Policies*, *5*(2), 74-87.

Shah, J.H. (1998) *Validity and credibility of public examinations in Pakistan*, Doctoral dissertation),

Retrieved from *http://eprints.hec.gov.pk/1020/*

Shavelson, R. J., & Webb, N. M. (1991) *Generalizability theory, A primer*. Newbury Park: Sage Publications.

Shirazi, M. J. H. (2004) *Analysis of examination system at university level in Pakistan*, Doctoral dissertation,

Retrieved from *http://eprints.hec.gov.pk/311/*

Starch, D., & Elliott, E. C. (1912) Reliability of grading high-school work in English. *School Review, 20*(7), 442-457.