Bulletin of Education and Research
August 2020, Vol. 42, No. 2 pp. 163-179

High Stake Testing: Factors Affecting Inter-rater Reliability in Scoring of Secondary School Examination

Sehar Rashid[*] and Nasir Mahmood[**]

Abstract

The study aimed to validate the factors that affect the inter-rater reliability of secondary school certificate (SSC) papers in high-stake testing. For this purpose, papers of Urdu and English of Board of Intermediate and Secondary Education (BISE) were selected. A survey method was used to collect marking on the same set of papers from each rater and their response to the questionnaire. The population of the study has comprised the raters for English and Urdu subjects. A sample of 98 raters was selected randomly from the list. Instruments of the study were solved papers of annual examination for both subjects and a questionnaire related to the factor that affects inter-rater reliability. According to consistency estimates of the inter-rater reliability approach, the Spearman correlation coefficient was applied to examine consistency in the scoring of raters. It was found that raters' training effect the inter-rater reliability in their scoring. However, other factors did not affect inter-rater reliability in scoring. Based on the findings of the study, it is recommended to ensure the participation of all raters in the training session.

Keywords: Inter-rater reliability, Consistency estimates of inter-rater reliability, Secondary School Examinations

[*]Department of Education, Virtual University of Pakistan. Email: sehar_rashid12@yahoo.com
[**]Professor and Dean Faculty of Education, Allama Iqbal Open University, Islamabad.
Email: mahsir1@yahoo.com

Introduction

The reliability of marking is a significant aspect to ensure quality control of the assessment procedure that affects a candidate's life chance. A major characteristic of reliability of marking is the inter-rater reliability. Inter-rater reliability is the extent to which a student obtains the same scores if different teachers scored the performance or rate the performance (Nitko, 1996). It is a quality indicator of marking reliability. It is useful to endorse the fairness of the criteria of marking and to uphold a clear understanding among raters. Consequently, it is significant to assess the inter-rater reliability in the scoring of raters to ensure marking reliability as well as the credibility of the high-stake testing.

Secondary school examination is a high stake testing for every candidate as a result of these examinations provide a chance to select the career of their own choice. Results of these exams takenby BISE also provide the baseline in merit selection for high secondary institutes. In such circumstances, it is important to maintain high inter-rater reliability in the scoring of raters so that the marks of any candidate should not depend on who marked the paper. Consequently, inter-rater reliability is crucial for high stake testing.

Numbers of research have been done in the nineteenth century with the concern of marking reliability which draws attention towards its importance. Research by Porterand Jelinek, (2011) found the range of inter-rater reliability from poor to moderate. It is generally considered that the subjectivity of the raters highly affects the essay type material. Therefore a measure of inter-rater reliability can be better assessed for the essay type material. Some of the researches emphasized the measure of inter-rater reliability for the essay type material of the papers. Rashid and Mahmood (2016) found moderate inter-rater reliability in the scoring of high-stake testing. This shows that concern of inter-rater reliability remains a subject need to be studied.

All earlier studies have generated further several studies globally with the concern to assess the inter-rater reliability which also identifies different factors that influence the inter-rater reliability. Those factors can be drawn as technical factors and personal factors.

Technical factors

The effect of equipment used for scoring of the examination during the assessment is said to be technical factors e.g. scoring scheme, training of examiners, etc.

Scoring criteria. The important stage for the scoring of examination by raters is the 'setting of standards' which leads to developing the rating system or scoring rubrics. The development of the scoring rubrics or rating system needs to be completed earlier to the marking session by the examination body. The practice of pre-defined marking criteria (rating system, answer keys, or scoring rubric) in the marking procedure is assumed to

lessen the subjectivity involved in rating of restricted response questions and extended type questions, accordingly increasing rater reliability (Moskal, 2000). Scoring rubrics answer this concern by formalizing the measuring scheme for separately score level. The explanations at each of the score levels are used to lead the assessment procedure (Moskal & Leydens, 2000). It illustrates that the progressive assessment procedure might be the result of an explicit scoring scheme for the marking of the papers.

The scoring scheme could be a prime device to attain high inter-rater reliability in the scoring of raters. The scoring scheme must be carefully developed as it is an important factor to attain consistency in the scoring of different examiners. The development of a scoring scheme is itself a critical step and needs an understanding of assessment objective in explicit terms. Saunders and Davis (1998) observed the construction and application of the scoring criteria for the scholar studies of management students and draw three conclusions; at first, the involvement of the examiners in the development of assessment criteria is useful to ensure that each of the examiners understands it well; secondly, there is need to debate on criteria from time to time to maintain the consistency; lastly, they emphasized the significance of flawless marking process and the impression that these processes require not to act as a restraint.

The introduction of the scoring method is a significant part of the scoring scheme. However, it is necessary to select an appropriate scoring method while developing the scoring scheme.

There are two main scoring methods among the several being in use currently that are, analytic scoring and holistic scoring. Wage (2009) concluded that the marks given by using the analytic marking were ordinarily a little higher than that of holistic scoring. Thus while developing the scoring scheme for marking papers by the raters, there must be a kind consideration about the scoring method suitable for particular items of the papers. It is understood that the development of the assessment criteria is not a one-shot procedure. To develop a clear, understandable, and precise assessment criterion, examination bodies have to plan and execute it sensibly and spotlessly with no or minimum error chance.

Training of examiners. It is significant to provide training to the examiners to guide them about the scoring scheme as the raters from the different areas might be varied in their achievement levels. There should be a scheduled procedure for the training of examiners that must be known to the examination bodies. Rudner (1992) proposed that to finest lessen examiners faults, examiners training plans should familiarize raters with the scoring procedure they supposed to follow, guarantee that raters comprehend the order of processes which have to essentially accomplished, then describe in what way the examiners ought to infer any normative data provided to them.

Training of examiners plays a substantial role in the scoring of exams. Shohamy, Gordon, and Kramer (1992) found that general consistency was greater for trained examiners as compared to the untrained examiners. Consequently, the training of examiners for a better understanding of the scoring process may have signed for the greater inter-rater reliability in the scoring of examiners.

The community of assessment practice. Another fundamental factor that has a direct concern with the assessment process is the community of assessment practice which may increase consistency in the scoring of examiners. Explicitly, consistent marking is hypothesized to be the result of an operative community of practice. The literature on the concept of a community of practice was initiated by the research of Lave and Wenger (1991). Wenger (1998) indicated that "*practice includes both the explicit and the tacit*" (p.47). The tacit knowledge is innate and commonly held. A community of practice is defined as "a group of people who share a concern, set of problems, or a passion about a topic, and who deepen their knowledge and expertise in this area by interacting on an ongoing basis" (Wenger, McDermott & Snyder, 2002).

Hall and Harding (2002) were accountable to devise the notion 'community of assessment practice' by the research aimed to enhance the consistent application of assessment criteria. Wolf (1995) contended that evaluator systems or conversation among raters is required for consistency. There are several research pieces of evidence to provision this dispute. A report by the Higher Education Quality Council (HEQC, 1997) on evaluation in higher education maintained that reliable assessment results among examiners are the results of communications time by time; the internalization of patterns, and comprehensive linkages. Orr and Nuttall (1983) found that in the examinations of the subject of English, it is the raters' consultations instead of the scoring criteria that are the central device for upholding consistency. Breland and Jones (1988) studied that high reliability in scoring is attained by examiners' work as a group (in a meeting situation) as compare if they work independently. So the importance of the community of assessment practice is not negotiable when the concern is related to the measure of inter-rater reliability.

Wenger (1998) claimed that as part of a community of assessment practice provides examiners a sense of possession of information and training and that it is uniform grading or involvement in taking decisions that maximally accelerate knowledge. Accordingly community of assessment practice motivates examiners by strengthening confidence in their scoring.

Ecclestone (2001) investigates the impact of a community of assessment practice in higher education and argued that only the scoring criteria and guidelines will not be enough to completely communicate the reliable standards unless the assessment staff is not socializing into the assessment community. It is concluded that to make the effective criteria a reliable one, there must be a community of assessment practice for examiners.

Personal factors

The personal traits of examiners that influence the scoring of the papers by examiners are said to be personal factors such as the experience of examiners, fatigue/ tiredness overtime, etc.

Experience of the examiner. The experience of the examiner is not an ignorable character. Moria (2003) concluded that an individual's feature establishes as essential in explaining reliability is the number of years of marking experience.

It is debatable that for the marking of the examinations for a particular subject, teaching experience is required or the marking experience or both. Royal-Dawson (2004) determined that the criteria of instruction practice could be calm to permit scoring. Ham (2001) concluded that mediator and rater experience was supplementary vital as compared to the teaching experience for reliability in the scoring of raters. Therefore it is likely that teaching experience and marking experience both are differentially important.

Tiredness/ Fatigue over time. Another vital trait of consistency in scoring is whether raters differ in their marking reliability over time. If so, an applicant's scores would be different conferring to time as a certain paper was scored. Few pieces of researchprovide evidence that the raters' fatigue over time affects their marking. Humphris and Kaney (2001) examined the concern of fatigue over time in raters and found slight proof of an orderly bias which might be inferred as so because of fatigue or tiredness.

Moira (1999) studied that an examiner who was reflected lenient initially in the marking period was occasionally reflected severe later on (or vice versa) and concluded that this could be affected by over-compensation for severity/leniency emphasized in the early checks. Moria, Massey, Baird, and Morrissy (2001) concluded that there were merely slight changes in the comparative leniency or severity of raters' overtime of the scoring session.

Thus the effect of fatigue over time to examiners on the scoring is significant to overcome for the attainment of high inter-rater reliability in the scoring.

Most of the studies in the literature exhibited low inter-rater reliability in the scoring of the examiners. Low inter-rater reliability is considered a key problem for test administrators, but good reason (Deboer, 2013). As the check of inter-rater reliability leads the administration toward progressive decisions for further or future proceedings of the assessment procedure.

The measure of inter-rater reliability has been researched globally since the last century. There are few pieces of researchconducted nationally to study the examination system of the country (Shah, 1998; Bashir, 2002; Shirazi, 2004; Kiani, 2004; Jaffri, 2006; Jilani, 2009) with the concern to analyze the validity, reliability, and effectiveness of the examination system. Rashid and Mahmood (2016) initiate the concern of inter-rater reliability and found moderate inter-rater reliability in the scoring of raters of high stake-testing. This provides a basis to validate the factor that affects inter-rater reliability in the scoring of high-stake testing.

This study was aimed to validate the effect of different factors on inter-rater reliability in SSC papers of high-stake testing. The objective of the study was to validate the effect of technical factors (scoring scheme, training of raters and community of assessment practice) and personal factors (marking experience and fatigue to raters) related to raters as factors that influence inter-rater reliability in the scoring of SSC papers of high stake testing. These objectives were attained by dealing with the following questions:

1. Is there any variance in the inter-rater reliability of SSC papers of BISE based on raters' training?
2. Is there any variation in the inter-rater reliability of SSC papers of BISE based on raters' participation in the community of assessment practice?
3. Is there any variance in the inter-rater reliability of SSC papers of BISE on the base of raters' marking experience?
4. Is there any variance in the inter-rater reliability of SSC papers of BISE based on raters' fatigue over time?

Method

Research design

The research design for the study was descriptive. A survey method was used to collect the data from a group of people that describe the aspects and characteristics of the population.

Population

The population was comprised of 539 raters for Urdu subject and 345 raters of English subject who participated in marking of relevant subject for the board of intermediate and secondary education (BISE) Lahore.

Sample of the study

A list of raters was collected from BISE Lahore ensuring the officials use that only for research purposes. A sample of 98 raters for both subjects was selected randomly. Consent on the telephonic conversation was taken from each rater to participate in the study before visiting them personally.

Instruments

Two research instruments were used to collect data from the selected sample which was

Questionnaire. It was developed according to the "rater's selection criteria for paper marking" and "guidelines for the paper rating/scoring" provided to the raters to rate/score papers by BISE Lahore. For this, related information and document were taken from the relevant office. The questionnaire also addressed the literature-based factors which affect inter-rater reliability that need to validate. It was comprised of the demographic profile of the raters and 26 statements related to the factors affecting inter-rater reliability. A dichotomous scale (yes or no) was used to get the responses from the sample of the study.

Table 1
Factor-wise distribution of the statements of the questionnaire

Factors	No. of the statement in the questionnaire
Features of training	15
Experience of raters	1
Scoring Scheme/Criteria	5
Fatigue over time	2
The community of assessment practice	1
Total	26

Validation of an instrument. The questionnaire was validated through repeated consultation with the supervisor.

Piloting and reliability of the questionnaire. The questionnaire was piloted to the 60 raters of both subjects during the scoring session in the BISE office. The reliability value of the questionnaire on collected data was 0.86 Cronbach's alpha, which is considered as a high-reliability value.

Solved papers. To collect this instrument, we conducted the annual papers of grade 10^{th} for the subjects of Urdu and English. These papers were conducted on the students of secondary level (grade 10^{th}) who appeared in the relevant session. The part of supply type items of the question paper for both subjects was not administered to the students as the subjectivity of the raters did not affect the marking of these items. The solved papers for both subjects consisted of restricted response items and extended response items of the

question papers. After conducting the papers, we have collected 4 papers for each subject (4 for Urdu and 4 for English). All these 8 papers were used as an instrument for the examiners of English and Urdu respectively.

Data analysis

Consistency estimates of inter-rater reliability. Consistency estimates of the inter-rater reliability approach were selected to measure inter-rater reliability in the scoring of raters. According to this approach, Spearman's correlation coefficient was selected to measure inter-rater reliability in the scoring of the raters. The consistency in the scoring of raters was benchmarked according to Landis and Koch (1977) benchmark levels.

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

Descriptive statistics. Along with the statistic technique to assess the inter-rater reliability in the scoring of the raters, some of the descriptive statistics like percentage, frequencies and mean value were applied.

As each category (trained raters and untrained raters etc.) was comprised of more than two raters, thus there were multiple numbers of comparisons for each category. After measuring inter-rater reliability for each category, minimum and maximum correlation value was recorded to measure variance inconsistency and, the mean of correlation values of the whole group was calculated to get an average of variance inconsistency for each category.

Findings and conclusions

Features of the scoring scheme provided to raters for the scoring of papers Figure 1 shows that more raters were that they were provided the scoring scheme for the marking of the papers, the scoring scheme was contained marking guidelines for the restricted response items, the scoring scheme was contained guidelines for the extended response items of the question paper, and, the scoring scheme was contained number-wise marking guidelines for extended-response items of the questions paper. On the other side, more raters disagreed that the scoring scheme was contained model answers for the extended response items of the question paper. Thus, it can be concluded that the scoring scheme provided to the raters for scoring purpose contained reasonable features as discussed in the literature.

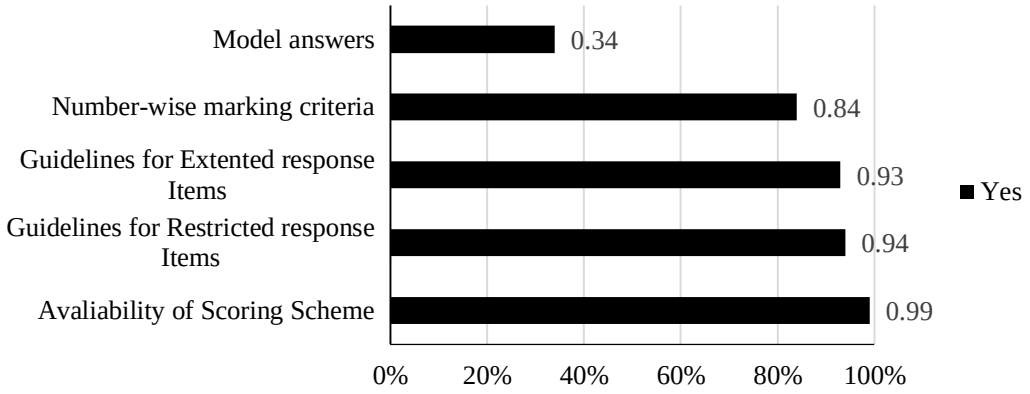

Figure 1.*Bar graph of rater's responses to the scoring scheme*

Variance in inter-rater reliability of scoring due to training of rater

For this research, those raters who attended the two-day workshop conducted by BISE Lahore for them to guide the marking session are said to be the trained raters. Figure 2 shows that more raters were trained.

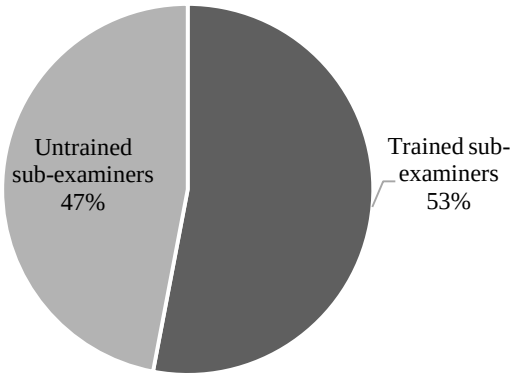

Figure 2. *Pie chart for the percentage of trained and untrained raters*

The trained raters respond to the statements of the questionnaire features of training. The results are given in figure 3.

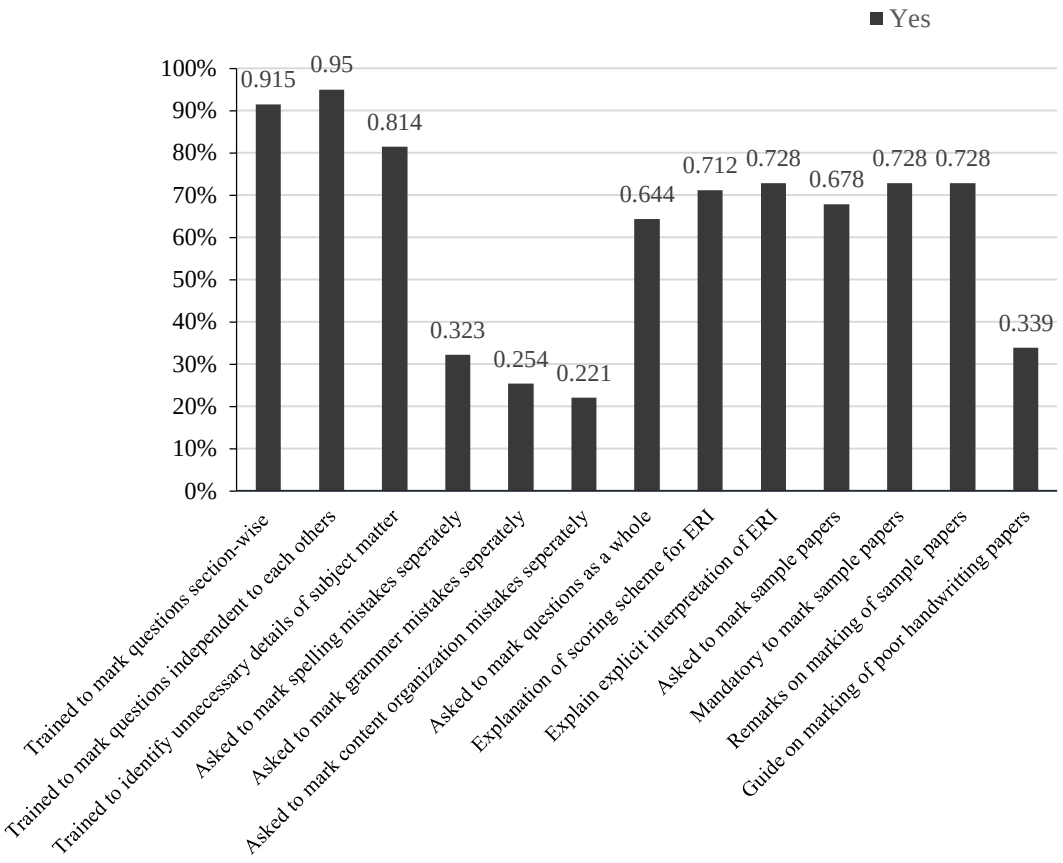

Figure 3.*Column graph of training of raters*

The scoring of the trained raters and untrained raters was measured to assess the effect of training of raters on the inter-rater reliability of the scoring of the papers. Table 3 reveals that the inter-rater reliability in the scoring of the trained raters exhibited less variance (substantial to almost perfect) as compare to the inter-rater reliability in the scoring of the untrained raters (moderate to almost perfect). The mean of 'r' shows substantial inter-rater reliability for untrained teachers and almost perfect inter-rater reliability for trained teachers.

Table 3
Comparison in inter-rater reliability of trained and untrained raters

Group	No. of raters	No. of comparisons	Variance inconsistency (r)	Mean (r')
Trained	52	760	0.614 to 0.991	0.862
Untrained	46	507	0.441 to 0.968	0.674

**Correlation is significant at 0.01 level (2-tail)
N=98

Variance in inter-rater reliability in scoring due to community of assessment practice

Table 4 reveals that the inter-rater reliability in the scoring of the raters who take part in the community of assessment practice exhibited less variance (moderate to almost perfect) as compare to the raters who do not take part in the community of assessment practice (fair to almost perfect). The mean of 'r' shows substantial inter-rater reliability for both groups.

Table 4
Comparison in inter-rater reliability of raters for a community of assessment practice

Group	No. of raters	No. of comparisons	Variance inconsistency (r)	Mean (r')
Participated	87	1993	0.443 to 0.991	0.717
Not participated	11	55	0.368 to 0.979	0.620

**Correlation is significant at 0.01 level (2-tail)
N=98

Variance in inter-rater reliability of scoring due to marking experience of raters

Table 5 reveals that there is much variance in inter-rater reliability (moderate to almost perfect) of raters based on their making experience. The mean of 'r' shows substantial inter-rater reliability in the scoring of the papers for all the three categories of the raters based on marking experience.

Table 5
Comparison in inter-rater reliability of raters based on rater's marking experiences

Marking experience	No. of raters	No. of comparisons	Variance inconsistency (r)	Mean (r')
≥ 3 years	33	495	0.398 to 0.953	0.693
4 to 10 years	39	708	0.313 to 0.952	0.655
10 < years	26	292	0.368 to 0.954	0.673

**Correlation is significant at 0.01 level (2-tail)
N=98

Variance in inter-rater reliability of scoring due to fatigue to raters

Raters' fatigue overtime was classified into two groups as fatigue over time due to the continuous marking for a longer time and the number of papers they mark at once.

Table 6 reveals that there is much variance (fair to almost perfect) in the interrater reliability of raters on the base of their time for marking. The mean of 'r' shows substantial inter-rater reliability in the scoring of the papers for all the 3 groups of the raters who can mark papers attentively up to 3 hours, 4 to 5 hours, and more than 5 hours continuously.

Table 6
Comparison among scoring based on rater's concentration span

Marking period	No. of raters	No. of comparisons	Variance inconsistency (r)	Mean (r')
1-3 hours	33	190	0.258 to 0.990	0.684
4-5 hours	39	289	0.267 to 0.934	0.657
5< hours	26	181	0.391 to 0.915	0.669

**Correlation is significant at 0.01 level (2-tail)
N=98

Table 7 reveals that there is much variance in inter-rater reliability (fair to almost perfect) of raters who can mark 1 to 37 and 38-42 papers in a day, and, for raters who mark 43 to 50 papers in a day (moderate to almost perfect). The mean of 'r' shows substantial inter-rater reliability in the scoring of the papers for all the three groups of the raters on the base of the number of papers they marked consecutively.

Table 7
Comparison among scoring of raters on the basis on no. of papers mark

No. of papers	No. of raters	No. of comparisons	Variance inconsistency (r)	Mean (r')
1-37	23	91	0.243 to 0.916	0.654
38-42	59	811	0.258 to 0.929	0.661
43-50	16	56	0.434 to 0. 962	0.704

**Correlation is significant at 0.01 level (2-tail)
N=98

Discussion

The study was aimed to validate the factors affecting inter-rater reliability in the scoring of SSC paper. Inter-rater reliability when more than one rater is involved in marking procedure. The scoring scheme is a vital component of the marking procedure, especially where raters are supposed to mark essay type items. In the study, raters reported that the scoring scheme was provided for the marking of the papers by BISE Lahore. That scoring scheme contained the basic elements to describe the marking of the different types of

items. So it might help decrease the subjectivity and increase the inter-rater reliability in the scoring of the raters. This finding is supported by the argument of Moskal (2000) that the use of a scoring scheme is assumed to lessen the subjectivity as well as increase the inter-rater reliability in the scoring of raters.

One more finding exhibited that an explicit marking scheme might not be the concern of experts who developed scoring scheme as some of the important guidelines about the making of the papers was missing or not described properly (model answers for the extended response items, number-wise scoring of the extended response items, etc.). It might hinder the evenness in the marking procedure. This finding negotiates the argument that explanation at each of the score level is used to lead the assessment process (Moskal & Laydens, 2000). Thus a trivially developed scoring scheme might be a threat to inter-rater reliability in the scoring of papers of BISE.

Variance in inter-rater reliability of the scoring of trained and untrained raters. Variance in inter-rater reliability in the scoring of rater was measured and findings of the study lead us to the conclusion that although there is a wide range of variance inconsistency of scoring for both groups of the raters even than the inter-rater reliability in the scoring of trained raters is less varied as compared to the inter-rater reliability of the untrained raters. This finding of a study is supported by the finding of the research by Shohamy, Gordon, and Kramer (1992).

On the other hand, the wide range of variance in inter-rater reliability for the scoring of trained raters may be caused by the unsuccessful training session or point out the non-productive training session. Alderson, Clapham, and Wall (1995) state that if training of the examiners is not helpful to have reliable marking then all other works done previously to have quality instruments for marking is of no yield. Thus the high variance in the inter-rater reliability in the scoring of trained rater has appeared to be considered all other measures taken to produce reliable results consciously by the BISE.

Variance in inter-rater reliability of scoring due tothe community of assessment practice. The findings of the present study suggest that the community of assessment practice is not guaranteed to high inter-rater reliability in the scoring of raters of BISE. This finding is contrary to the findings of other studies (HEQC, 1997; Orr & Nuttall, 1983; Breland & Jones, 1988; Ecclestone, 2001).

The results of the study can be explained in terms of other factors that might cause variance in inter-rater reliability in the scoring of the raters. The key point which can explain the findings of the study is that all the raters of the BISE were not trained to make the scoring scheme understandable to the raters. Another finding of the present study revealed the wide range of variance in the scoring of trained raters which lead us to

the conclusion that raters might not understand the scoring scheme clearly. If the raters did not understand the scoring scheme or were not trained for the marking session then it is no matter how much assessor network or discussion made by raters, the variance in the inter-rater reliability might difficult to overcome.

Variance in inter-rater reliability of scoring caused by marking experience of raters. The finding of the study revealed that the marking experience of the raters was not a determinant to high inter-rater reliability in the scoring of raters of BISE. The finding is contrary to the findings of other studies (Ham, 2001) that can be described in terms of environmental conditions provided to the raters while marking sessions by BISE. Asrater gain their marking experience by scoring the papers of BISE only which is of great chance, inthe same situations identified by this research (trivially constructed scoring scheme, less productive training, and unbeneficial community of assessment practice). Marking experience gain by raters in all these conditions might not be as productive as it should be. Thus the finding under discussion is not a peculiar result of the study for the raters of BISE.

Variance in inter-rater reliability of scoring due to fatigue over time to raters. As far as the factor of fatigue overtime is concerned, the findings of the study revealed that the factor of fatigue over time to raters was not the cause of variance in inter-rater reliability for the scoring of the papers as there is a wide range variance already exist in inter-rater reliability of all the raters irrespective to the factor of fatigue over time. This finding is contrary to the finding of another study (Humphris & Kaney, 2001). The finding of the study can be described in terms of the daily routine of the raters during marking session in BISE. Raters had to join marking sessions after a tough routine of school as raters should be a schoolteacher as per policy. So the daily routine of the raters while marking sessions observed by us was that they attended their schools in the morning (for 6 hours) and then they directly went to the BISE offices in the afternoon (for 3 to 6 hours) to perform their duty as a rater. Thus it can be concluded that the raters took the factor of fatigue with them to the BISE offices to perform their rater duties. It is ahead stricken point that if a rater starts scoring of paper in fatigue condition then how we can expect consistency in the scoring of the raters.

Another thought-provoking thing was that most of the teachers tried to achieve the task set by the BISE officials (one rater can mark up to 50 papers in a day) after reaching the BISE offices (for 3 to 6 hours) to perform their duty. This created haste in the rater to fulfill the target in 3 to 6 hours instead of 11 hours which was set for a whole day (from 9am to 8pm) by BISE. Thus, the factor of haste intermingles with the factor of fatigue at the start of marking is a major threat to the inter-rater reliability in the scoring by the raters.

On the basis of conclusions and discussion, it is recommended that there is a need to make it mandatory for all the raters to attend the training session to become eligible for the marking of the papers. The training session should cover all the aspects of marking procedure that supposed to practice by the raters. It is also recommended that the rule of the BISE Lahore that one can mark up to 50 papers in a day or we can say in 11 working hours (from 9am to 8pm) should be reconsidered according to the hours one is going to spend to perform his/her duty as a rater to eliminate the factor of haste and fatigue overtime to raters.

References

Alderson, J. C., Clepham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

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

Breland, H. M., & Jones, R. J. (1988). *Remote scoring of essays.* (ETS RR 88-4) New York:

Deboer, F. (2013, September 13). *Do we want perfect inter-rater reliability?* Retrieved September 28, 2013, from http://fredrikdeboer.com/2013/09/13/do-we-want-perfect-inter-rater-reliability/

Ecclestone, K. (2001). "I know a 2:1 when I see it": Understanding degree standards in program franchised to colleges. *Journal of Further & Higher Education, 25*(4), 301-313. DOI:10.1080/03098770126527

Hall, K., & Harding, A. (2002). Level descriptions and teacher assessment. *Educational Research, 44*(1), 1-16. DOI:10.1080/0013188011008107 1

Ham, V. (2001). Maintaining National Standards in Standards-Based Assessment: The New Zealand Experience. The *University of Leeds.* British Educational Research Association.

Higher Education Quality Council (1997). *Assessment in Higher Education and the role of' Graduateness'*. London: H.E.Q.C., Graduate Standards Program.

Humphris, G. M., & Kaney, S. (2001). Examiner fatigue in communication skills. *Medical Education, 35*, 444-449.

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

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 Punjab, based on Solo Taxonomy*, Doctoral dissertation, Retrieved from http://prr.hec.gov.pk/Thesis/774S.pdf

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

Lave, J., & Wenger, E. (1991). *Situated learning legitimate peripheral participation.* Cambridge: Cambridge University Press.

Moira, A. P. (1999). *Office Review evaluation*. A Research Report, RC12.

Moria, A. P. (2003). *Examination background and the effect on marking reliability.* RC218: AQA Research Report.

Moria, A. P., Massey, C., Baird, J., & Morrissy, M. (2001). *Marking Consistency over time.* AQA Research Report: RC/129.

Moskal, B. M. (2000). Scoring rubric: what, when, and how? *Practical Assessment, Research, and Evaluation, 7*(3), 1-7.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation*, 7(10).

Nitko, A. J. (1996). Educational assessment of students (2nd ed.). United States of America:Printice-Hall, Inc.

Orr, L. & Nuttall, D. (1983). *Determining standards in the proposed single system of examining at 16+*. London: Schools Council.

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.

Rashid, S. & Mahmood, N. (2016). High-Stake Testing in Punjab: Inter-rater Reliability in the Scoring of Secondary School Certificate (SSC) Examination. *Journal of Research and Reflections in Education. 10*(2), 156-168.

Royal-Dawson, L. (2004). *Is teaching experience a necessary condition for markers of Key Stage 3 English?* AQA Research Report, RC261.

Rudner, L. M. (1992). Reducing errors due to the use of judges. *Practical Assessment, Research & Evaluation, 3*(3). Retrieved from http://pareonline.net/getvn.asp?v=3&n=3

Saunders, M. N., & Davis, S. N. (1998). The use of assessment criteria to ensure consistency of marking: Some implications for good practice. *Quality Assurance in Education, 6*(3), 162-171. DOI:10.1108/09684889810220465

Shah, J.H. (1998).*Validity and credibility of public examinations in Pakistan* (Doctoral dissertation). Retrieved from *http://eprints.hec.gov.pk/1020/*

Shirazi, M. J. H. (2004). *Analysis of examination system at the university level in Pakistan* (Doctoral dissertation). Retrieved from *http://eprints.hec.gov.pk/311/*

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27-33.

Wang, P. (2009, September). The inter-rater reliability in scoring composition. *English Language Teaching, 2*(3).

Wenger, E. (1998). *Communities of practice learning meaning and identity.* Cambridge: Cambridge University Press.

Wenger, E. C., McDermott, R., & Snyder, W. C. (2002). *Cultivating communities of practice: A guide to managing knowledge.* USA: Harvard Business School Press, Cambridge.

Wolf, A. (1995). *Competence-based assessment.* Buckingham: Open University Press.