Voyage Journal of Educational Studies (VJES) **Vol. 4 Issue 1**
ISSN (Online): 2790-7171, ISSN (Print): 2790-7163 **January to March 2024**

Reliability Issues in Standardized Assessment: A Study of the SSC English Examinations in Khyber Pakhtunkhwa

Mehran Ali
Research Scholar, M.Phil. English (Linguistics), City University of Science and Information Technology, Peshawar
Corresponding
mehranalam89@gmail.com

Aiman Naeem
Research Scholar, MPhil English (Linguistics), City University of Science and Information Technology, Peshawar

Alia Rasool
PhD Scholar, English (Linguistics), Department of English, Islamia College, Peshawar

Muhammad Umer
Assistant Professor of Linguistics, Department of English, Islamia College Peshawar

Abstract

This research attempted to examine the reliability issues in the standardized assessment of the SSC English examinations in Khyber Pakhtunkhwa. Discrepancies in students' achieved score on the examinations and their actual scores raise questions about the grading of the said English examinations. This study employed a survey with open-ended questions. The data were collected from students who scored 90% or above marks on the examinations through random sampling. The data were analyzed both qualitatively and quantitatively. The preliminary findings revealed a significant discrepancy between students' grades and their actual writing proficiency as identified through inter-rater grading. The study revealed a noticeable depreciation in inter-rater grading compared to the Board grades, thereby creating doubts about the reliability of the grading practices of the standardized English assessment. It is suggested that paper grading methods and techniques of board examinations are revisited.

Keywords: *Reliability in Assessment, Grading Reliability, English Language Assessment, Interrater Reliability, SSC Examination, KP.*

Introduction

Secondary School Examination (SSC) is one of the most important examinations in Pakistan taken by high school students who seek admission in reputable institutions to continue their intermediate level studies. Most students after their SSC Examination enter colleges with poor English language writing proficiency which raises questions regarding paper grading and assessment practices in Board exams. To have problems with writing is highly alarming as it is an important language skill that students' academic life highly relies on. Poor writing proficiency leads to poor performance in exams and assignments. To investigate the issue, an exploration of English language assessment practices, particularly grading practices, is highly needed. Once the testing and evaluation are reliable, authentic, and valid, it will aid in the achievement of all of these objectives and will yield positive effects. Normally, language tests are designed in such a way as to reflect an individual's language ability and proficiency (Alderson & Wall, 1993).

In language teaching, assessment refers to a variety of approaches and procedures, mostly tests, used by language assessment experts to evaluate, measure, and verify students' language proficiency (Hughes, 2003). Testing provides crucial diagnostic information for educational groups, identifies instructional needs, and recommends instructional interventions for teachers. For parents and students, testing information provides a measure of individual success, academic strengths, and weaknesses that aid them in making decisions (Herman, 1992). Hence, language assessment and testing have a great role in learning a language, as testing is used by nearly all educational systems due to its impact on how and what the teachers teach and students learn, and, therefore require a solid and reliable framework for evaluating students to prepare them for future challenges (Ramirez, Schofer, & Meyer, 2018). The efficiency of teaching and the impact produced can be measured by looking into the depth of students' learning by conducting valid tests having reliable grades.

This preliminary research therefore aims to investigate the reliability of English paper scores in the SSC Examination of BISE Mardan, to see if it reflects the actual language proficiency of students. The reason for choosing this area of research is the poor reputation of the BISE examinations in Pakistan (Greaney & Hasan, 1998). The current state of assessment in Pakistan is plagued by several significant issues, such as corruption and cheating during the examination process and the quality of examinations and the marking process which is often compromised

(Everett & Burdett, 2017). Consequently, the results obtained from these assessments prove to be a poor predictor of a student's actual abilities (Rind & Mari, 2019). Addressing these issues is crucial to establishing a fair, transparent, and effective assessment system that genuinely reflects the academic achievements and capabilities of students across the country. Therefore, this research tries to highlight if the board exams, with a specific focus on BISE Mardan, are reliable in their marking. The findings will inform the policymakers about the grading practices and reliability of English language assessment and will contribute to the existent body of knowledge in the domain of English language assessment particularly about grade reliability.

Literature Review

In education, the concept of assessment refers to a variety of approaches and procedures used by educators to evaluate, measure, and verify students' academic performance, learning progress, skills enhancement, and educational needs (Brown & Abeywickrama, 2018). Assessments also detect individual student deficits and strengths, allowing educators to provide specific academic support, educational management, or social endeavors. For administrators and school planners, test results provide information and highlight areas of curricular strength and constraint. Testing provides crucial diagnostic information for educational groups, identifies instructional needs, and recommends instructional interventions for teachers (Brown, Race, & Smith, 1996). For parents and students, testing information provides a measure of individual success, academic strengths, and weaknesses that aids them in enhancing their learning and making decisions based on students' progress (Vergis & Hardy, 2009). Reliability, validity, and washback are the features of a good assessment. Once the testing and evaluation are reliable, authentic, and valid, it will aid in the achievement of all the objectives and will cast a positive washback.

The term "reliability" refers to the capacity to believe that a grade or judgment was given based on actual performance rather than some insignificant factor or scoring context (Herman, 1992). Numerous elements have an impact on a student's test score and consequently on its reliability. The degree to which a test score is independent of the following four factors: the day and hour of the test, questions or difficulties, or the rater who rates the test is referred to as reliability. Multiple test editions with distinct questions or problems aimed at examining the same common skills or areas of knowledge, and different raters rating the test takers' responses are all examples of reliability (Livingston, 2018)

Reliability is important because it checks whether a test score is significant, if it doesn't show, at least roughly; how the test taker would have done if the test had been given on a different day, how the test taker's responses would have been assessed by a different set of raters if the test taker had been given a different set of questions or problems to measure the same general skills or knowledge (Fulcher & Davidson, 2007). There are several types of reliability, each of which refers to a distinct level of consistency. The consistency of test takers' performance across multiple editions of the test is known as "alternate-forms reliability." The consistency of the scores given by different raters to the same answers (essays, performance samples, etc.) is known as “interrater reliability.” Stability (or "test-retest reliability") refers to the consistency of test takers' performance throughout days or times of testing. (Livingston, 2018, p. 7).

Alternate-Forms Reliability and Internal Consistency

The consistency of test takers' scores across different editions of the exam, comprising different questions or problems measuring the same categories of knowledge or skills at the same difficulty level, is known as alternate-forms reliability (Livingston, 2018, p. 13). It clarifies the doubt to what extent a student performs well in one edition and then in another edition of the same test. It applies to any test that exists in more than one edition. Even if another edition is unavailable, the result can still be generalized to other questions that were not part of the concerned edition. It provides details on an inconsistency caused over which the test designers have some control. They can improve the alternate-forms accuracy of the scores by making the test longer and including more questions or problems (ibid). Internal consistency refers to the consistency of test takers' responses to different questions or problems on the same edition of the test. It answers the question, "How well do test takers who perform well on one topic also perform well on other questions?" Internal consistency will be high if all the questions test the same knowledge and ability. If the questions were about different things, it would be low. Is it important to have internal reliability? There are two reasons for presenting internal reliability parameters. The use of data at hand—the test takers' responses on a single test edition—to generate these values. Second, when using an assumption that is usually close to the truth, internal consistency values are a good estimate of alternate-forms reliability statistics. Internal consistency statistics are produced for this reason, not because internal consistency is significant, but because they usually provide a reasonable estimate of alternate-forms reliability.

Inter-rater Reliability

It is the level of agreement between the scores generated by different raters who scored identical responses (Fulcher & Davidson, 2007, p.15). Only the choice of raters is included as a key cause of difference. It's conceivable that the test taker's score would have been different if another set of raters had assessed those responses. This issue can be addressed using inter-rater reliability statistics. If a test taker took two separate editions of an exam that were both scored by raters, it is improbable that the same raters would score both editions. Even if the pool of raters was the same for both editions, the same rater may not be assigned an individual test both times. The alternative types of test reliability will cover both the selection of raters and the selection of questions or issues as possible reasons for inconsistency (i.e., measurement error). For high alternate form reliability, the inconsistency effects of the above sources should be kept in control. Even if test takers' performance is consistent, their scores on the two editions of the test will not be consistent if the scoring is inconsistent (inter-rater reliability is low). Strong inter-rater reliability does not imply high alternate-forms reliability, but low inter-rater reliability does. (Livingston, 2018, p. 16).

Factors Influencing Reliability

Multiple factors can influence a person's test score when he or she takes it. The most important aspect that has influenced a test taker's score is the skill, knowledge, and ability that the test is designed to measure. On the other hand, the test taker's score will frequently be based on other types of knowledge and talents that the test does not assess. A collection of abilities known as "test wasteness" is another factor (Mousavi, 2009, p. 804). One such ability is the capacity to effectively use testing time. Another is knowing when and how to guess on a multiple-choice test. Two things that can affect a test score are the test taker's alertness and focus on the day of the test. When taking tests, as in many other endeavors, most people perform better on certain days than others. If you take an exam while alert and able to concentrate, your score is more likely to be higher than if you take it while fatigued or distracted. The questions or concerns that most tests provide to the test taker are not the only ones that may have been included. Different versions of the test describe various questions or problems assessing the same type of knowledge or skills. You have probably had the good fortune to take an exam that asked you questions about what you had learned at some point during your education. A test taker with strong abilities in the abilities assessed by the test will perform well on any edition of the test—but not equally well on each version (Livingston, 2018).

The test score is also influenced by the test rater. The outcome of an essay writing test given in a class and then assessed by two raters can be different. It's possible that one rater's preferred style and method of presenting ideas in an essay drew him in and caused him to give it a high rating, while another rater was completely frustrated by the essay and gave it a low rating (Brown & Abeywickrama, 2018, p.30). So, the chance factor has its influence on both cases. These are the factors that have the potential to raise or lower a test score. There is little further that can be done to mitigate the impact of these circumstances other than to advise learners to relax and be more alert on exam day. This can also be resolved by dividing the test into sections, although in most circumstances, this is not doable. By presenting students with the entire course and contents before the test, the chances of undesired questions in the test can be lowered. To reduce the odds of differing ratings, a rating instruction list should be presented to the rater, and a scoring standard should be established, with the rater being required to follow the instructions (Ahmad & Politt, 2011).

In short, it is impossible to declare the test scores of test takers to be pure and free from every kind of chance factor. Chances factors influence test takers' performance in a test. Test takers are humans; they can have issues, can have problems focusing, can have deficiencies, and could not be perfect in all aspects. The raters are too human. Everyone has different tastes and knowledge. These are the chance factors that normally influence the test score. This influence can be minimized through different ways but cannot be completely eradicated. The test score that comes after the minimization of the influence of the chance factors is the reliable score. Reliability is the test result that is free from every sort of influencing factor. Reliability is consistency. When there is consistency in test scores of test takers in different editions of tests of the same skill and difficulty level is known as alternate-forms reliability. It is to measure the test takers' strength of performance in different editions once they perform well in one edition. This can be improved by adding more questions to the existing edition and increasing its length. The rating also has an influence over reliability. When the same responses are rated by different raters independently, then it is called interrater reliability. When the two editions of the test are rated by different rater then the score will be different. It is to check the inconsistency in the score there and to eradicate it and make the score more reliable. An assessment or score of the test takers will not be valid when it is not declared reliable. For this reason, reliability must be ensured to produce a valid and authentic assessment.

Research Methodology

The study used a survey as a research strategy and a questionnaire as a research tool. An open-ended questionnaire was designed so that the respondent could freely express themselves. A random sampling technique was followed where the respondents were selected purely at random from the delimited population to avoid biases and prejudice. However, data collection was delimited to those students of BISE Mardan who achieved more than 90% marks in the SSC Exam. Furthermore, the research has used both quantitative and qualitative modes of data analysis as per the nature of the study. The collected data was transcribed and tabulated. Data transcription was cross verified to ensure entries as per students' original responses. Two tables were designed: a demographic table containing the participants' name, gender, board roll no, institution attended, total marks obtained out of 1100 in the SSC Exam, and marks obtained in English Paper; another table was a comparison table where the marks obtained by the participants in English were correlated with the marks awarded by four experts of English. An additional table was also made to trace the frequency of spelling mistakes per answer by the participants. The tables can be found in the Data analysis and tabulation section.

Checking Questionnaires from Experts

The filled questionnaires were checked for spelling and grammar mistakes from three graduates of Peshawar University with having Master's in English. The checkers were requested to mark the paper out of a total of 72 marks. The questionnaires contained 9 questions, so every question carried 9 marks which computationally equal 72 marks. The researcher awarded every participant with 3 marks just for the reason to make it equal to the actual board marks of English, which are 75.

Expert A

Expert is MA in English from the University of Peshawar in the year 2006. He has fifteen years of teaching English experience currently working as a Certified Teacher and teaches English in a government sector high school under the Elementary and Secondary Education Department, KP. He has experience of marking English papers in BISE, Mardan.

Expert B (Subject Specialist in English)

Expert holds MS in English from Northern University, Nowshera. He has 17 years of teaching English experience which has quite polished his teaching, checking, and assessment skills. He is currently working as a Subject Specialist in English in a government sector higher secondary school, under the Elementary and Secondary Education Department, KP.

Expert C (Subject Specialist in English)

Expert C has done his MA in English from the University of Peshawar. He has 20 years of teaching English currently working as a Subject Specialist in English in a government sector higher secondary school under the Elementary and Secondary Education Department, KP.

All the experts were given a total of 10 transcribed questionnaires on January 26, 2021, for analysis and grading. The expert analyzed and graded the questionnaires and returned them to the researcher on February 15, 2021. As research ethics, the identities of the experts have been kept confidential.

Data Analysis

The following tables contain data analysis of 10 students who took 95% marks in the English paper of the SSC Examination of Mardan board. Each table contains the student's demographic information and a thorough analysis of the student's writing. After analyzing student's writing manually by checking it from field experts, the nature and frequency of mistakes along with its graph of comparison with total words is therefore reported. Similarly, a graphical comparison of interpreters marks and students' actual grades has also been provided. Students' opinions regarding paper attempts have also been presented in the last row of each section to obtain an in-depth, rich, and holistic picture of the situation.

Table 1

The following table comprises of five rows. The first and second row contains table's description and students' demographic information, marks obtained in Board Examination, and in English paper. Likewise, the third row contains total words written by students along with nature and number of mistakes committed. Similarly, the fourth row compares marks awarded by Board Grader and inter-rater grading to same student along with percentage, graphical presentation, and correlation of awarded grades. The last row contains students' opinion regarding obtaining good marks.

Table 1

Student Name	Roll number	School status	School Name	MO in SSC out of 1100 (1100)	MO in English out of 75 (100%)
Student S	N/A	Private	XYZ	1074 (97%)	73(97%)

Mistakes in Writing of the Student

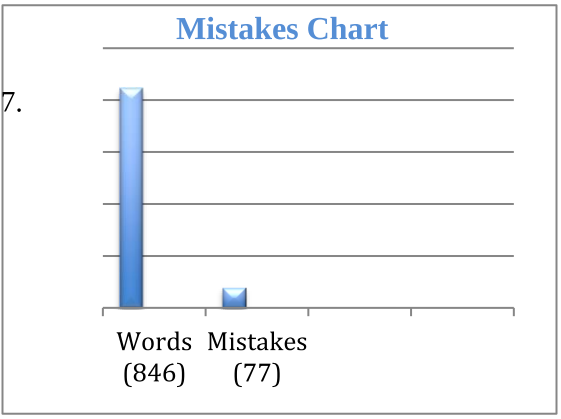
Out of a total of 846 words, the student has made 77. Common mistakes. The percentage of mistakes is therefore 9.01% out of 100%. The mistakes Reflects the student's deficiency in following. Areas.

Poor Punctuations Literacy.

Low frequency words mistakes.

Faulty usage of Modal auxiliaries.

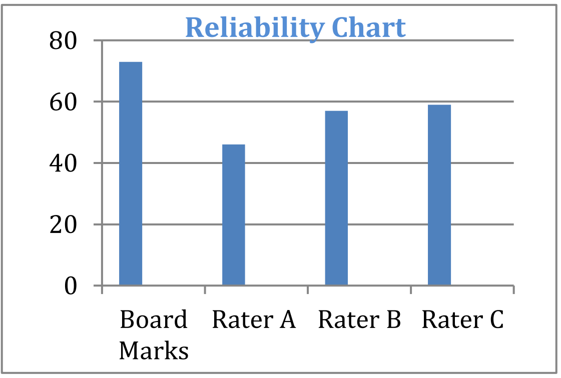
No literacy of using definite Articles.



Establishing Grades Reliability Through Interrater Reliability:

The student has achieved 73(97%) marks in Board. Exam. However, Rater A has awarded 46 (61%) marks, while Rater B 57 (76%)
And Rater C 59 (78%) marks to the student.
to the student.

The marking shows rating inconsistency across Various raters. However, the cumulative Percentage of all the interpreter is still low 72% Then that of the rater at Board 93%, thereby reflecting unreliability at board level.



Opinion of the Student regarding getting good marks in board:
Use of highlighter, colors, margin lines, and headings

Table 2

The following table comprises of grades given by different raters to the writing task of another student. The graph shows that student R has made some minor grammatical mistakes in the writing. However, the marking by different raters gives a surprisingly different result. The student has received 96% marks in the board exam, while the same student has been given an accumulative 77% marks by three different raters, thus raising a question mark on the difference between

the student's actual language proficiency and his/her board marks.

Table 2

Student Name	Roll number	School status	School Name	MO in SSC out of 1100 (1100)	MO in English out of 75 (100%)
Student R	N/A	Private	XYZ	1049 (95%)	72 (96%)

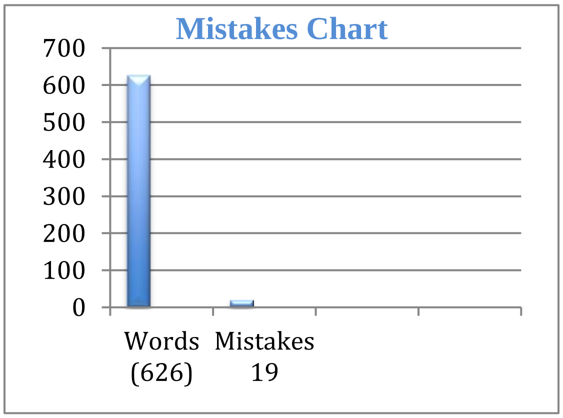
Mistakes in Writing of the Student

Out of total 626 words, the student has made 19. Common mistakes. The percentage of mistakes is therefore 3.03% out of 100%. The student is. quite good at expression. However, the mistakes Reflects the student's deficiency in following. Areas.

Poor understanding of Affixes

Verb/Tense level mistakes.

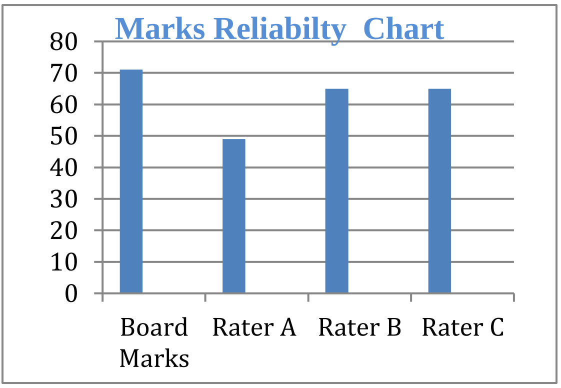
Subject Verb Agreement issue.



Establishing Grades Reliability Through Interrater Reliability:

The student has achieved 71(94%) marks in Board. Exam. However, Rater A has awarded 45 (60%) marks, while Rater B 65 (86%) And Rater C 65 (88%) marks to the student.

The marking shows rating inconsistency across Various raters. However, the cumulative Percentage of all the interpreter is still low 77% Then that of the rater at Board 94%, thereby reflecting unreliability at board level.



Opinion of the Student regarding getting good marks in board:
Lining, Good presentation, good handwriting, Flow charts, Headings.

Table 3

The following table discusses the results and graphical representation of the scores achieved by Student T. The student committed some minor grammatical mistakes, but overall, there is clarity in style and expression. However, inter-rater reliability of the student shows a striking difference between his/her board marks and the individual rating by three raters. The percentage difference can be seen in the following table.

Table 3

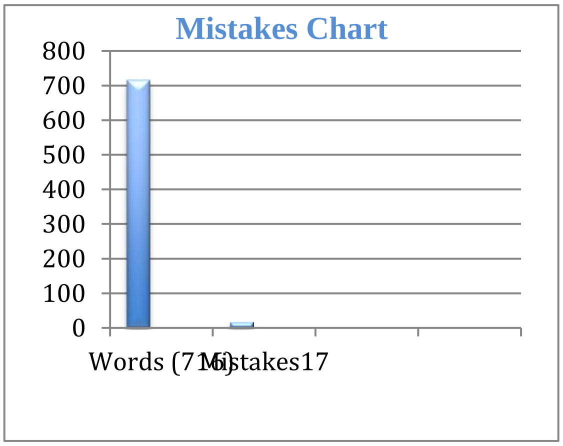
Student Name	Roll number	School status	School Name	MO in SSC out of 1100 (1100)	MO in English out of 75 (100%)
Student T	N/A	Private	XYZ	1049 (97%)	73 (97%)

Mistakes in Writing of the Student

Out of total 716 words, the student has made 17. Common mistakes. The percentage of mistakes is therefore 2.37% out of 100%. The student has good writing style and clarity in expressions. However, the mistakes reflect the students. deficiency in following areas:

Use of contractions in Academic writing.

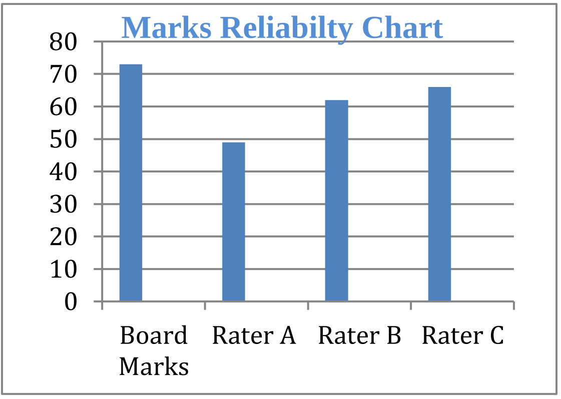
Subject verb Agreement.



Establishing Grades Reliability Through Interrater Reliability:

The student has achieved 73(97%) marks in Board. Exam. However, Rater A has awarded 49 (65%) marks, while Rater B 62 (82%)
And Rater C 66 (88%) marks to the student.
to the student.

The marking shows rating inconsistency across Various raters. However, the cumulative Percentage of all the interpreter still low 78% than that of the rater at Board 97%, thereby reflecting Unreliability at board level.



Opinion of the Student regarding getting good marks in board:
Blue and Black markers, Headings, Impressive Paper, Lengthy writing, and underlining.

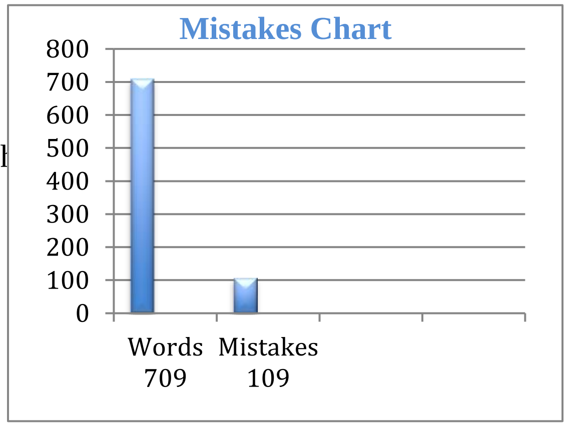
Table 4
The following table comprises of five rows, with the results displayed for Student P. It includes the marks awarded to the student by different raters, and his opinion regarding obtaining good marks. Details are given in the following table.

Table 4

Student Name	Roll number	School status	School Name	MO in SSC out of 1100 (1100)	MO in English out of 75 (100%)
Student P	N/A	Private	XYZ	1049 (95%)	72 (96%)

Mistakes in Writing of the Student

Out of total 709 words, the student has made. 107 Common mistakes. The percentage of Mistakes are therefore 15.09% out of 100%. Th Reflects the student's deficiency in following. Areas.



Poor Academic Literacy.

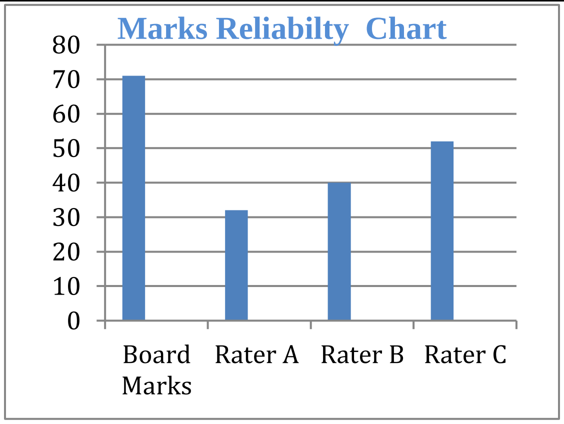
Low frequency words mistakes.

Subject Verb Agreement and tense issue.

No literacy of using definite Articles.

Establishing Grades Reliability Through Interrater Reliability:

The student has achieved 71(94%) marks in Board Exam. However, Rater A has awarded. 32 (61%) marks, while Rater B 40 (76%) And Rater C 52 (78%) marks to the student. to the student.



The marking shows rating inconsistency Across Various raters. However, the Cumulative Percentage of all the interpreter is still low at 55%. Then that of the rater at Board 94%, thereby reflecting unreliability at board level.

Opinion of the Student regarding getting good marks in board:

Headings, neat writing, Margin lines, and use of different colors.

Table 5

Table E enlists the percentage, graphical presentation, and correlation of grades for Student C. The results show that the student has committed grammatical mistakes, including misappropriate use of words, tense issues, and subject-verb agreement. To our surprise, the student achieved 95% marks in the board exam. The same student has been graded an accumulative 73% by the raters, which leads us to believe that the board marks awarded generously. Further details are given in the table below.

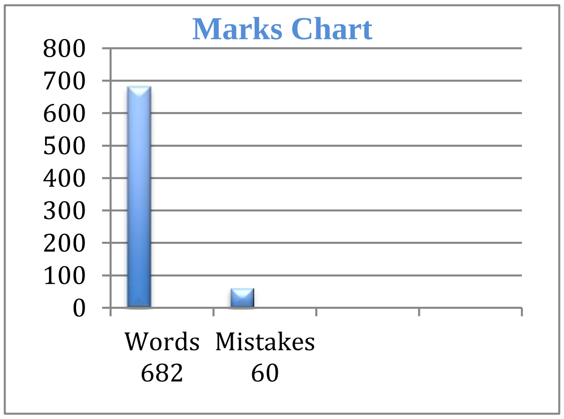
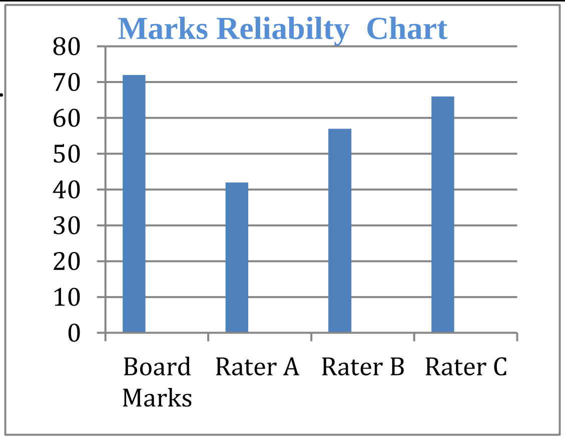
<table>
<tr><th>Student Name</th><th>Roll number</th><th>School status</th><th>School Name</th><th>MO in SSC out of 1100 (1100)</th><th>MO in English out of 75 (100%)</th></tr>
<tr><td>Student C</td><td>113504</td><td>Private</td><td>XYZ</td><td>1029 (95%)</td><td>74 (96%)</td></tr>
<tr><td colspan="6">Mistakes in Writing of the Student
Out of total 682 words, the student has made 60. Common mistakes. The percentage of mistakes is therefore 15.09% out of 100%. The mistakes Reflects the student’s deficiency in following. Areas.
Malapropism.
Tense issue.
Subject Verb Agreement issue.
Poor usage of definite Articles.
Poor Punctuation Literacy
 </td></tr>
<tr><td colspan="6">Establishing Grades Reliability Through Interrater Reliability:
The student has achieved 74(94%) marks in Board. Exam. However, Rater A has awarded 42 (61%) marks, while Rater B 57 (76%)
And Rater C 66 (78%) marks to the student. to the student.
The marking shows rating inconsistency across Various raters. However, the cumulative Percentage of all the interpreter are still low 73% than that of the rater at Board 94%, thereby reflecting unreliability at board level.
 </td></tr>
<tr><td colspan="6">Opinion of the Student regarding getting good marks in board:
Headings, Paper Charming, and underlining.</td></tr>
</table>

Table 6

The following table consists of results shown for Student G. The student has achieved 95% marks in his board exams. However, data analysis reveals that the student is not even able to make meaningful sentences, especially compound and complex sentences. Similarly, the student committed mistakes related to tenses and pronouns. The interrater percentage of the marks given to the student is 67%, which puts a question mark on the grades awarded by the board. Details can be seen in the following table.

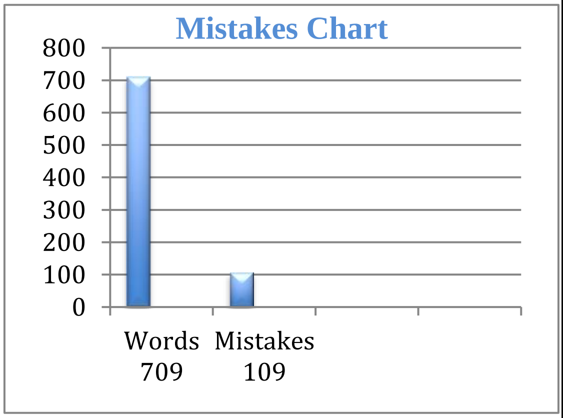
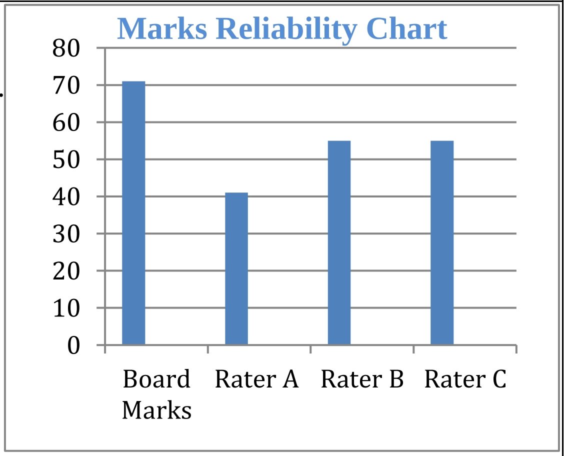
<table>
<tr><th>Student Name</th><th>Roll numbe r</th><th>School status</th><th>School Name</th><th>MO in SSC out of 1100 (1100)</th><th>MO in English out of 75 (100%)</th></tr>
<tr><td>Student G</td><td>N/A</td><td>Private</td><td>XYZ</td><td>1027 (95%)</td><td>71 (96%)</td></tr>
<tr><td colspan="6"><b>Mistakes in Writing of the Student</b>
Out of total 673 words, the student has made 69. Common mistakes. The percentage of mistakes is therefore 10.25% out of 100%. The mistakes Reflects the student’s deficiency in following. Areas.
Tense and Verb issue.
Second person pronoun usage issue.
Most of the sentences make no sense.

 </td></tr>
<tr><td colspan="6"><b>Establishing Grades Reliability Through Interrater Reliability:</b>
The student has achieved 71(94%) marks in Board. Exam. However, Rater A has awarded 41 (61%) marks, while Rater B 55 (76%) And Rater C 55 (78%) marks to the student. to the student.
The marking shows rating inconsistency across Various raters. However, the cumulative Percentage of all the interpreter are still low 67% Then that of the rater at Board 94%, thereby reflecting unreliability at board level.

 </td></tr>
<tr><td colspan="6"><b>Opinion of the Student regarding getting good marks in board:</b>
Coloring, Highlighting and Headings.</td></tr>
</table>

Table 7
The following table demonstrates the comparative marking and feedback on the writing skills of Student O. This table also shows the same result for the students, who are getting a higher percentage in board exams while getting comparatively lower when marked by individual raters.

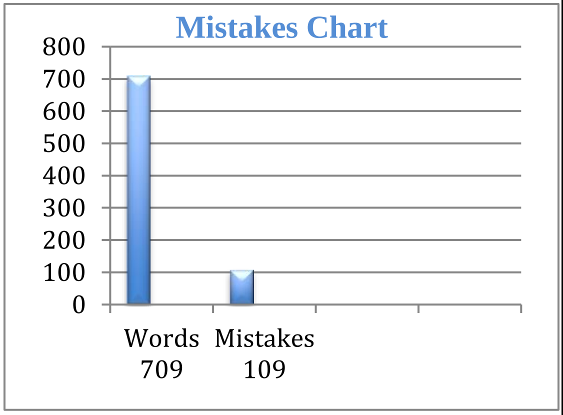
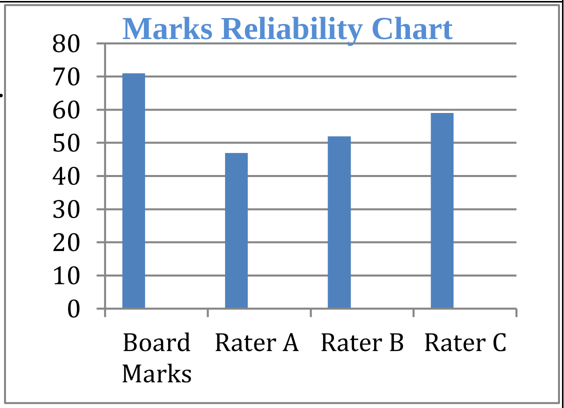
<table>
<tr><td>Student Name</td><td>Roll number</td><td>School status</td><td>School Name</td><td>MO in SSC out of 1100 (1100)</td><td>MO in English out of 75 (100%)</td></tr>
<tr><td>Student O</td><td>95576</td><td>Private</td><td>XYZ</td><td>1074 (97%)</td><td>71 (94%)</td></tr>
<tr><td colspan="4"><b>Mistakes in Writing of the Student</b>
Out of total 643 words, the student has made 55. Common mistakes. The percentage of mistakes is therefore 8.55% out of 100%. The mistakes Reflects the student's deficiency in following. Areas.
Very long and complex sentences.
Low frequency words mistakes.
Subject Verb Agreement and tense issue.
Capitalization issue.
Wrong use of Punctuations.</td><td colspan="2"> </td></tr>
<tr><td colspan="4"><b>Establishing Grades Reliability Through Interrater Reliability:</b>
The student has achieved 71(94%) marks in Board. Exam. However, Rater A has awarded 47 (62%) marks, while Rater B 52 (69%) And Rater C 59 (78%) marks to the student. BBB to the student.
The marking shows rating inconsistency across Various raters. However, the cumulative Percentage of all the interpreter are still low 70% Then that of the rater at Board 94%, thereby reflecting unreliability at board level.</td><td colspan="2"> </td></tr>
<tr><td colspan="6"><b>Opinion of the Student regarding getting good marks in board:</b>
Conceptual and to the point writing, no use of colors and highlighter.</td></tr>
</table>

Table 8

The following table comprises of the details of the total words written by students along with nature and number of mistakes committed. Similarly, it highlights the difference between marks awarded by the Board and the other three raters. Details of the comparison can be seen in the following figure.

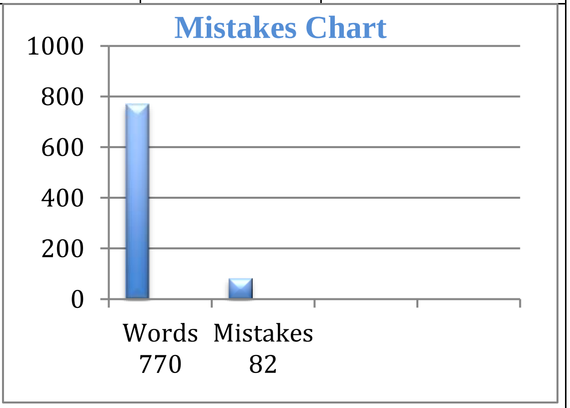
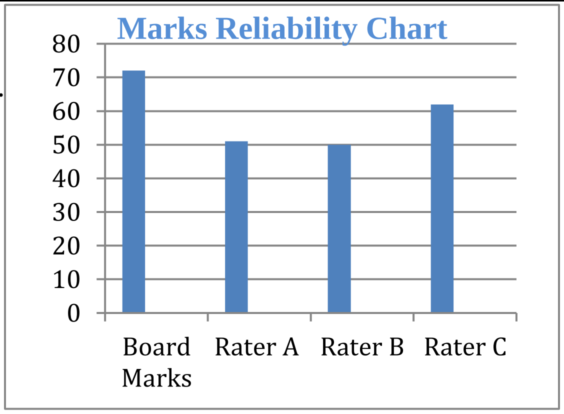
<table>
<tr><td>Student Name</td><td>Roll number</td><td>School status</td><td>School Name</td><td>MO in SSC out of 1100 (1100)</td><td>MO in English out of 75 (100%)</td></tr>
<tr><td>Student Y</td><td>95576</td><td>Private</td><td>XYZ</td><td>1074 (97%)</td><td>72 (96%)</td></tr>
<tr><td colspan="4">Mistakes in Writing of the Student
Out of total 770 words, the student has made 82. Common mistakes. The percentage of mistakes is therefore 10.64% out of 100%. The mistakes Reflects the student’s deficiency in following. Areas.
Concordance issue.
Poor capitalization Literacy.
Fragmented sentences.
Wrong use of Punctuations.</td><td colspan="2"> </td></tr>
<tr><td colspan="4">Establishing Grades Reliability Through Interrater Reliability:
The student has achieved 72(96%) marks in Board. Exam. However, Rater A has awarded 51 (62%) marks, while Rater B 50 (69%) And Rater C 62 (78%) marks to the student. to the student.
The marking shows rating inconsistency across Various raters. However, the cumulative Percentage of all the interpreter are still low 72% Then that of the rater at Board 96%, thereby reflecting unreliability at board level.</td><td colspan="2"> </td></tr>
<tr><td colspan="6">Opinion of the Student regarding getting good marks in board:
Margins, Proper heading and diagrams, and neat writing.</td></tr>
</table>

Table 9

The following table shows the results of scores for Student U. The data reveals that the student got a higher percentage of marks in board exams as compared to his actual marking by independent checkers. It can be assumed that the student committing grammatical mistakes cannot achieve 90+%. However, this student has been awarded quite a good score, which affects the reliability of these exams. Details of mistakes are given in the figure below.

Student Name	Roll numbe r	School status	School Name	MO in SSC out of 1100 (1100)	MO in English out of 75 (100%)
Student U	95583	Private	XYZ	1010 (91%)	71 (94%)

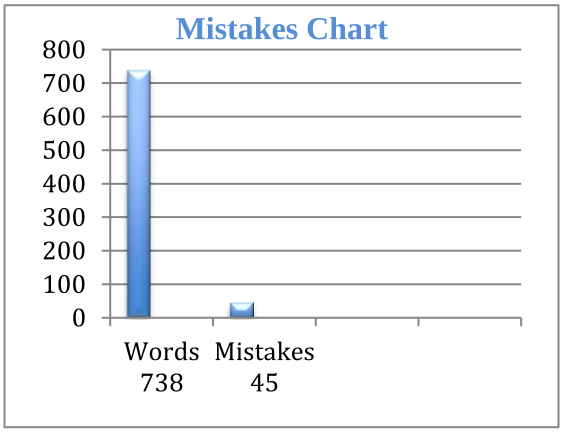
Mistakes in Writing of the Student

Out of total 738 words, the student has made 45. Common mistakes. The percentage of mistakes is therefore 6.09% out of 100%. The mistakes Reflects the student's deficiency in following. Areas.

A lot of contractions in Academic writing.

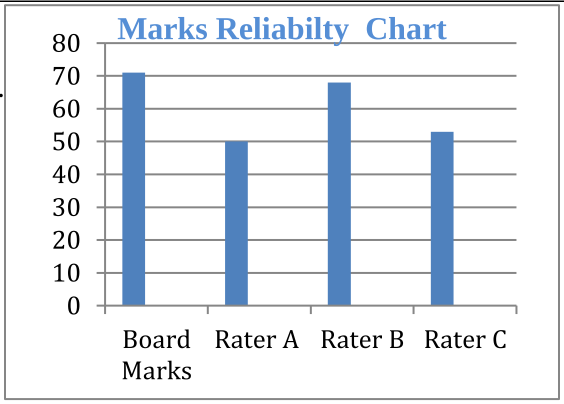
Poor Modal auxiliary usage Literacy.

Poor Definite Articles use Literacy.



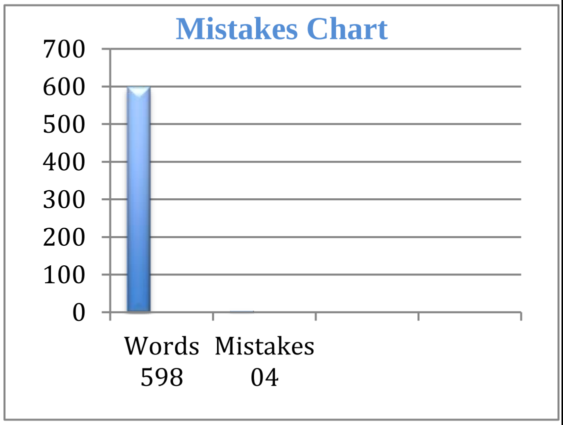
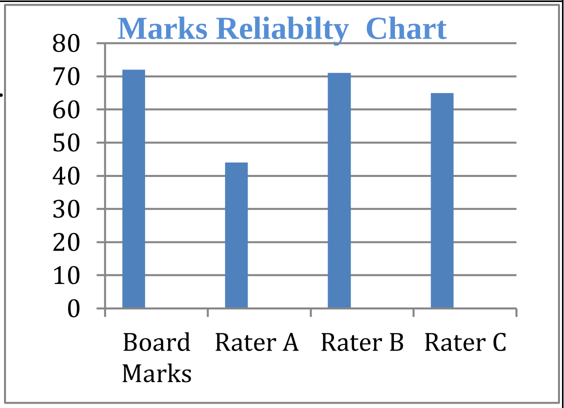
Establishing Grades Reliability Through Interrater Reliability:
The student has achieved 71(94%) marks in Board. Exam. However, Rater A has awarded 50 (66%) marks, while Rater B 68 (90%)
And Rater C 53 (70%) marks to the student. to the student.

The marking shows rating inconsistency across Various raters. However, the cumulative Percentage of all the interpreter are still low 76% Then that of the rater at Board 94%, thereby reflecting unreliability at board level.



Opinion of the Student regarding getting good marks in board:
First attempt easy questions, preferring answers that were best memorized and good handwriting.

Table 10
The last table presents the comparative marking and comments on the writing skills of Student W. This table also shows the same result for the students, who are getting a higher percentage in board exams while getting comparatively lower when marked by individual raters.

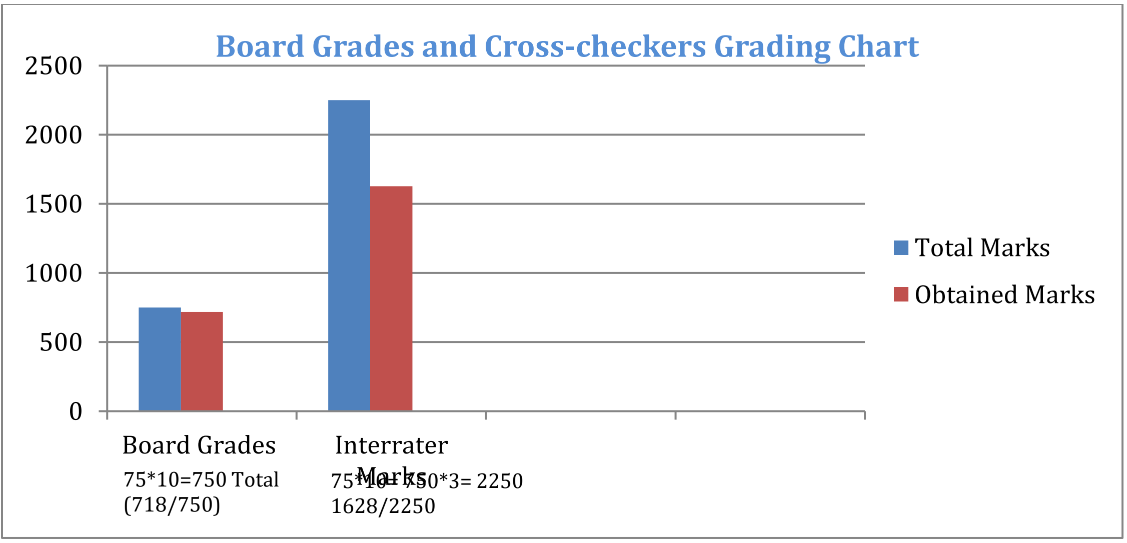
<table>
<tr><td>Student Name</td><td>Roll number</td><td>School status</td><td>School Name</td><td>MO in SSC out of 1100 (1100)</td><td>MO in English out of 75 (100%)</td></tr>
<tr><td>Student W</td><td>95576</td><td>Private</td><td>XYZ</td><td>1074 (97%)</td><td>72 (96%)</td></tr>
<tr><td colspan="4">Mistakes in Writing of the Student
Out of total 598 words, the student has made 4. Common mistakes. The percentage of mistakes is therefore 0.06% out of 100%. The mistakes Reflects the student’s deficiency in following. Areas.
Compound word usage issue.
Low frequency word spelling mistake</td><td colspan="2"> </td></tr>
<tr><td colspan="4">Establishing Grades Reliability Through Interrater Reliability:
The student has achieved 71(94%) marks in Board. Exam. However, Rater A has awarded 44 (58%) marks, while Rater B 71 (94%)
And Rater C 65 (86%) marks to the student. to the student.
The marking shows rating inconsistency across Various raters. However, the cumulative Percentage of all the interpreter are still low 80% Then that of the rater at Board 94%, thereby reflecting unreliability at board level.</td><td colspan="2"> </td></tr>
<tr><td colspan="6">Excellent presentation, good handwriting, and diagrams.</td></tr>
</table>

Board Grades and Cross-checkers Grading Chart
This chart shows the comparison of marks that are awarded to the students by the board and by individual raters/Subject Specialists who marked the answers of students from the filled questionnaires. The raters marked their paper in the same manner i.e. out of 75 as the Board exams English papers policy. A total of 10 students who secured 95% marks in English were

selected as samples for analysis. Out of a total of 750 (100%) Marks (75×10= 750), the board awarded 718 (96%) marks to the students. On the other hand, the three raters marked the questionnaires out of 75 (100%) with a total average of 225(100%) (75×3=225). Out of a total of 2,250 (100%) (225×10=2,250), the student secured 1628 (72%) marks. The grades awarded by the Board make 96%, while the average percentage of marks awarded by interpreter is 72% which is drastically lower than that of Board marks.

The marking of cross-checkers seems to be more reliable in comparison with Board Grades, because how can a person with technical mistakes in writing (as evident from the tables above) score 96% in a written exam? So, with reasonable confidence, it can be deduced that the Board Grades of English paper is somewhat unreliable and invalid as the statistics derived from the actual performance of the students put a question mark on Board Grades reliability.

Figure 1



Discussion and Conclusion

To investigate the relationship between students' English language writing proficiency with the reliability of test grades, a total of ten students with 95% marks in English was chosen as a sample for evaluation. A different version of the language test was created in the form of questionnaires to test the language proficiency of the students. To ensure reliability, factors affecting reliability were minimized because the term "reliability" refers to the capacity to judge grades based on actual performance rather than luck draw (Herman, 1992). So, the test scores were made independent of the following four factors: the day and hour of the test, questions or difficulties, or the rater reliability that affects reliability (Livingston, 2018). The students were made free from examination hall anxiety, time management, stress, and other negative factors. They were given a free hand in writing answers to questions. Similarly, all three raters were guided by the researcher to avoid discrepancies to a greater extent. The Cross-checkers rated the students' papers in the same way that the Board Grader graded their English papers, i.e. out of 75, as per board policy Because to reduce the odds of differing ratings, a rating instruction list should be presented to the rater, and a scoring standard should be established, with the rater being required to follow the instructions (Livingston, 2018).

After cross-grading questionnaires and avoiding all influencing factors, the students were supposed to have scored much better than their scores in the Board examination, but it was contrary to expectations. As per the grade's reliability chart in data analysis, the Board had awarded the student 718 (96%) out of a total of 75 *10 = 750 (100%) marks. However, three cross-checkers, on the other hand, scored the questionnaires out of 75 with a total average of 750*3=2250 (100%). The student received 1628 (72%) marks out of a total of 2,250 (100%). The grades awarded by the Board to the students were 96%, while the average percentage of marks awarded by inter-raters is 72% which is lower than that of Board grades. There is a depreciation of 24% between the cross-checkers and Board awarded grades.

The marking of cross-checkers seems to be more reliable than that of the board, because they have reported a significant number of mistakes each student committed. Each student's mistakes were quantified out of the total words written. The reliability of grading on board must therefore be questioned as to how students with technical mistakes in writing can (as evident from the tables in the analysis portion) score 96% in a written exam. Multiple reasons could be attributed to the unreliability of grading at the board level. The students either have memorized the lessons or there might have certain problems— mood swings, frustration, lack of proper training—with the grader who checks papers on the board. Furthermore, an interview with a grader from the board foreshadowed the main reason behind the unreliability of grading at the board level may be a financial one. For each paper, a grader is paid some good money. So, the more papers the checker checks, the more money he gets in return. For awarding high grades, the grader is normally not asked about as no student complains about having good marks. Moreover, to award someone low grades, the checker must check his paper for all possible mistakes which consumes the time of the checker and consequently affects the amount of money he makes. Conclusively, whether the problem is of memorization and reproduction, or the problem is with the checker at the board, the grading is unreliable—as a reliable test must always be valid—as evident from the grading of cross-checkers as their marking is in proximity with each other while that of Board grades show a deviation of 24%.

Recommendations for Policy Makers

Based on the research findings discussed, the following are some of the recommendations for policymakers and the Board officials.

- Existent policies regarding paper checking must be revisited.
- Checkers on board must be given proper training.
- A special committee must be established that can ensure fair and reliable grading.
- Tests must be designed in such a way to promote high-order learning.
- Validity of the test must be ensured by making significant changes in test design with the introduction of more real and life-like questions that could promote language learning.
- SLO based assessment may please be introduced.
- Reliable grading and inter-rater reliability at the marking level should be ensured.
- Graders before grading questions on the Board should be properly trained.
- A proper strategy should be devised to decrease chance factors and luck draw in grading.

Conclusion

To sum up, for matching students' grades with their English language writing proficiency, 10 students with 95% marks in English subjects were selected. Upon analysis of the data, there were major mistakes in word-level and sentence-level major mistakes in their writings that three of the cross-raters identified. Their writing did not reflect their actual English language writing proficiency as they committed a lot of common mistakes in their writing. The grades awarded to students on the board were therefore proved unreliable, as no one among the cross-checkers awarded them the marks equivalent to their board grades. However, there was 24% depreciation in the grading. The test marks, therefore, proved unreliable and invalid as the students did not perform well in writing sound answers in questionnaires that was a kind of different version of their board paper with the same marks. The study reveals that the Higher Secondary reading and writing tests are far below the satisfactory level in terms of reliability as they have a deficiency in both test reliability and scorer reliability. Therefore, the scores produced by this test cannot be considered reliable indicators of the test takers' reading and writing abilities. Based on the findings, some recommendations have been made for the improvement of the overall reliability of paper/test design and marking. Further research on classroom observation and syllabus design is suggested to investigate the impact of the same examination on teaching and learning practices.

References

Ahmed, S. & Rao, C. (2012). Examination washback effect: Syllabus, teaching methodology and learner communicative competence. *Journal of Education and Practice, 3(15): 173 - 183.*

Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice,* 18(3), 259-278.

Alderson, J.C, & Hamp, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing,* 13(3): 280-297.

Alderson, J.C., & Wall, D. (1993). Does wash back exist? *Applied Linguistics*, 14(2): 115-129.

Alwi, S. K. K., Rauf, M. B., & Soomro, S. (2016). Effects of cross and same age peer tutoring on reading attitudes of primary school students. *The Sindh University Journal of Education-SUJE*, *45*(1)

Bachman, L. (1990). *Fundamental considerations in language testing*. New York, NY: Oxford University Press.

Brown, H. D., & Abeywickrama, P. (2018). *Language Assessment: Principles and Classroom Practices*. 3rd ed., Pearson Education.

Burdett, N. and Everett, H. (2017). The impact of an examination board in Pakistan on student outcomes. *Conference paper, RISE, National Foundation for Educational Research.*

Cheng, L. (1998). Impact of public English examination change on students' perceptions and attitudes toward their English learning. *Studies in Educational Evaluation*, 24(3): 279-301.

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. United Kingdom: *SAGE Publications Ltd*

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment: An advanced Resource Book*. Routledge: New York

Greaney, V., & Hasan, P. (1998). 'Public examinations in Pakistan: a system in need of reform.' In: Hoodhood, P. (Ed) Education and the State: *50 Years of Pakistan. Karachi: Oxford University Press.*

Hughes, A. (2003). Testing for language teachers. Cambridge: *Cambridge University Press*

Herman, J. L. (1992.). A Practical Guide to Alternative Assessment. Alexandria.*: Office of Educational Research and Improvement (ED), Washington, DC.*

Kirkpatrick, R., & Gyem, K. (2012). Washback Effects of the New English Assessment System on Secondary Schools in Bhutan. *Language Testing in Asia*, 2(4), 5. https://doi.org/10.1186/2229-0443-2-4-5

Livingston, S. A. (2018). *Test Reliability—Basic Concepts.* Princeton, New Jersey: Test reliability—Basic concepts (*Research Memorandum No. RM-18-01), Princeton, NJ: Educational Testing Service.*

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed.) (pp. 13-103). *New York: Macmillan.*

Mathew, R. (2012). Understanding Washback: A Case Study of a New Exam in India'. In C. Tribble (Ed.), *Managing Change in English Language Teaching: Lessons from Experience,* (pp. 193-200). London, UK: British Council.

Mousavi, S. A. (2009). An encyclopedic dictionary of language testing (4th ed.). *Tehran, Iran: Rahnama Publications.*

Pearson, I. (1988). Tests as Levers for Change. In D. Chamberlain & R. J. Baumgardner (Eds.), *ESP in the Classroom: Practice and Evaluation. ELT Documents Volume 128* (pp. 98-107). London: Modern English Publications

Rehman, F. U., Rauf, M. B., & Alwi, S. K. K. (2018). Factors Effecting English Learning At Secondary School LEVEL: A Case Of Quetta. *New Horizons*, *12*(1), 113-150

Ramirez, F., Schofer, E., & Meyer, J. (2018). International Tests, National Assessments, and Educational Development (1970–2012). *Comparative Education Review, 62(000), 000-000.* https://doi.org/10.1086/698326

Rind, I. A., & Mari, M. A. (2019). Analyzing the impact of external examination on teaching and learning of English at the secondary level education. *Cogent Education, 6 (1).*

Sehar, S., Alwi, S. K. K., & Shaiq, M. Potential Benefits of Bilingual Teaching in Learning History

Saif, S. (2006). Aiming for positive wash back: A case study of international teaching assistants. *Language Testing*, 23(1): 1 -34.

Salihi, H. (2012). The washback effect of the Iranian universities entrance exam; teacher insights. *Journal of Language Studies*, 12(2): 609-628

Soomro, A. H., & Shah, S. Z. A. (2016). Effects of Washback on High School Teachers of English. *International Journal of English and Education*, 5(2), 201-210.

Soomro, M. N., & Memon, N. (2016*).* Investigation of Teacher as an Inducing Factor of Washback in Pakistan. *Language in India, 16(5), 185-198.*

Vergis, A. Hardy, K. (2009). Principles of Assessment: A Primer for Medical Educators in the Clinical Years. *The Internet Journal of Medical Education*, 1(1), 1-9.

Williams, M. & Burden, R. (1997). Psychology for language teachers: A social constructivist Approach. *Cambridge, England: Cambridge University Press.*