Journal of Education & Humanities Research (JEHR)
Institute of Education & Research (IER), University of Balochistan, Quetta-Pakistan
Volume: 18, Issue-II, 2024; ISSN:2415-2366 (Print) 2710-2971 (Online)
Email: *jehr@um.uob.edu.pk*
URL: *http://web.uob.edu.pk/uob/Journals/jehr/jehr.php*

"Exploring the Reliability of Current Summative Assessment Practices in English at the Secondary Level in Azad Jammu and Kashmir"

Batool Atta: *Assistant Professor, Institute of Education, University of Azad Jammu & Kashmir, Muzaffarabad, Pakistan*
Muhammad Altaf: *Research Scholar, Institute of Education, University of Azad Jammu & Kashmir, Muzaffarabad, Pakista*
Naveed Sarwar: *Assistant Professor, Department of English, University of Azad Jammu & Kashmir, Muzaffarabad, Pakistan*

Received: 12th November, 2024
Accepted: 30th November, 2024
Published: 31st December, 2024

KEY WORDS

Academic Integrity, Schools, Visually Impaired Students, Parents, Special Education

ABSTRACT

Reliability is a much important aspect of assessment activity. This study aims to determine what practices are carried out to maintain, ensure and improve the reliability of the certification examination in English at the secondary level in Azad Jammu and Kashmir (Pakistan). There is only one intermediate and secondary education board in Azad Jammu and Kashmir. All the sub-examiners (test graders), approximately 300 in number, were taken as a population of the study. A sample of 100 sub-examiners was drawn. The researchers applied a survey design and used a self-structured questionnaire as a data collection tool. Data were analysed through simple statistical measures, i.e., frequencies and percentages. The study results showed that examiners who mark the answer scripts are well-qualified and experienced but not particularly trained in evaluation. Some risks to the reliability of grading were also found. This study implies that immediate measures to improve the reliability of summative assessment systems, such as training examiners, must be taken.

Introduction

Assessment practices in the context of the SSC certification examination of AJKBISE have received limited research attention, despite the crucial requirement for assessments to be both valid and reliable. This study aims to investigate the strategies employed by sub-examiners to enhance the reliability of grading English answer scripts. Even at higher education levels, validity and reliability issues persist in assessment practices.

The Azad Jammu and Kashmir Board of Intermediate and Secondary Education (AJKBISE) Mirpur holds exclusive responsibility for conducting secondary level certification examinations in Azad Jammu and Kashmir, Pakistan. Established in 1973 through an ordinance, the board appoints head examiners and sub-examiners who grade bundles of answer scripts. Previously, Azad Jammu and Kashmir fell under the jurisdiction of the Board of Intermediate and Secondary Education, Lahore. However, with the establishment of AJKBISE, it gained autonomy. The inaugural examination conducted by AJKBISE took place in 1974, with 6,161 candidates appearing for the SSC and HSSC exams. The board adheres to rules and regulations aligned with those of the Lahore Board, ensuring educational standards parity with Punjab (AJKBISE, 2017).

Assessment serves as a vital link connecting content, teaching-learning practices, and outcomes, providing essential evidence for evaluating student progress. It also plays a pivotal role in facilitating learning itself. Assessments inform teachers, parents, and other stakeholders about students' achievements and progress (Teachers' Guide to Assessment, 2014). Despite careful planning and implementation of instructional strategies, it is impossible to predict with certainty what students have learned. Therefore, assessment becomes the means through which educators can ascertain the effectiveness of their instructional practices in achieving desired learning outcomes (William, 2013).

Reliability is a critical aspect of assessment, reflecting the consistency and stability of results generated by an assessment tool (Phelan & Wren, 2006). Various factors, such as the temporary physical and psychological states of test takers and environmental conditions in the testing site, can influence test performance (Carr, 2011). Subjective scoring in assessments can introduce inconsistencies in grading due to variations in experience, training, and perspectives among graders, thereby affecting the reliability of the test (Carr, 2011).

Grading constructed response test items, particularly in language tests, presents challenges when aiming for reliable results. Evaluating test takers' responses involves assessing multiple dimensions, including word choice, sentence structure, spelling, grammar, organization of ideas, coherence, and cohesion (Malone, 2017). However, there are concerns regarding the reliability of the examination process. Some examiners may prioritize securing grading jobs over contributing to educational quality improvement, potentially leading to a lack of efficiency in identifying mistakes in answer keys provided by examination boards (Kyani, 2011). Maintaining 100% validity and reliability in subjective test items, which often involve lengthy and detailed questions requiring extensive discussions, is a significant challenge (Murphy, 2006).

Assessment challenges are further amplified when the assessment subject is a second or foreign language. Abbara (2004) suggests that classroom tests in English as a foreign language often lack test-retest reliability due to construction flaws and the involvement of irrelevant or non-professional evaluators. Limited time for scoring may lead examiners to delegate the task to acquaintances, resulting in unreliable results when multiple people grade the same test. The validity of a test is

Corresponding Author's Email: batool.atta@ajku.edu.pk

partially dependent on its reliability (Ghazali, 2016), and factors such as test-retest reliability, internal consistency, and inter-rater reliability contribute to test validity (Joseph et al., 2020). Standardized tests generally exhibit the highest level of validity and reliability.

The competence of teachers significantly influences the quality of tests, and untrained teachers may struggle to ensure validity and reliability. Tests prepared and scored by less competent teachers tend to be less valid and reliable compared to those created and assessed by well-educated and trained teachers (Bartman et al., 2007). Untrained teachers may face difficulties in constructing valid test items and demonstrating satisfactory scoring of subjective test items (Shepard, 2009). Even at higher education levels, maintaining complete validity and reliability in subjective test items remains challenging. However, employing strategies such as dividing information into parts can contribute to improving reliability.

Grading tests is a responsible task that requires competent individuals with subject expertise who understand the significance of the test and the impact of the results on various stakeholders (Thissen, 2001). Irresponsible grading can render a test useless, wasting valuable time and resources (Thissen, 2001). Objective tests are generally easier to grade compared to subjective tests, as they save time on test construction and demonstrate high reliability (Kuramoto & Koizumi, 2018). However, even in objective scoring, human errors can occur, highlighting the importance of multiple graders and a clear answer key to minimize mistakes (Kuramoto & Koizumi, 2018).

Research suggests measures to enhance the reliability of grading, such as clear guidelines and grading frameworks should be provided to graders to ensure consistent standards are applied throughout the process (Fives & Barnes, 2013). Scoring rubrics can be employed, either with multiple graders or a single grader evaluating multiple test takers, to promote greater grading consistency (Fives & Barnes, 2013). The grading process itself impacts the quality of a test, as improper grading can lead to misleading results and hinder the effectiveness of the teaching-learning process (Gitomer et al., 2019). To improve the reliability of grading subjective test items, Dorgans and Cook (2020) suggest grading all responses to a specific essay at one time. If both writing quality and essay content need to be assessed, separate grades should be assigned before combining them. Using two graders for each essay and averaging their grades helps ensure consistency. Additionally, providing comments and correcting errors on test papers enhances feedback for students.

Lockwood et al. (2020) propose developing a scoring scheme that lists significant facts or theories to enhance the rigor of grading subjective categories. By assigning grades to these elements and incorporating them into the scoring scheme, the analytic scoring approach reduces halo effects and leniency errors, making it preferable to global scoring methods. While achieving complete validity and reliability in scoring is challenging, employing strategies can enhance it to an acceptable level. Murphy (2006) suggests dividing the necessary information or characteristics required to answer a question into distinct parts and assigning separate grades to each part.

This concise literature review accentuates the importance of assessment in the evaluation of student learning outcomes. The significance of reliability in assessments is highlighted, along with an exploration of various influencing factors. The challenges associated with grading constructed response test items are acknowledged, underscoring the necessity for implementing strategies such as analytic scoring, employing multiple graders, and utilizing technological advancements in assessment. Emphasis is

Corresponding Author's Email: batool.atta@ajku.edu.pk

placed on ensuring the professionalism and expertise of examiners. The importance of clear guidelines, scoring rubrics, and objective grading methods is underscored as crucial measures for enhancing reliability. These steps contribute to the attainment of accurate test results and support effective teaching and learning processes. In conclusion, this review underscores the significance of reliability in assessments and emphasizes the importance of implementing measures to enhance the evaluation process.

Methodology

The study employed a descriptive survey design. The researchers surveyed a selected sub-examiners of English who grade answer scripts. The population for this study comprises all English sub-examiners frequently working for the Board of Intermediate and Secondary Education, Mirpur. Due to the unknown and varying number of sub-examiners in different sessions/examinations, the researchers opted for convenience sampling and selected a sample of 100 sub-examiners.

A questionnaire based on research review was developed and utilized to collect as the primary tool. The questionnaire consisted of two sections. Section A included 11 questions that gathered demographic and other relevant information, which required rating scales. Section B comprised 24 statement-based questions, and respondents were asked to indicate their level of agreement using a five-point rating scale. The questions addressed various aspects, such as the examiners' eagerness to perform the evaluation job, efficiency in identifying mistakes in the provided answer key, strategies employed to maintain equality in grading, involvement of other evaluators, scoring quality, and satisfaction with remuneration.

The collected data from the questionnaires were analyzed by calculating the frequency and percentage of responses for each question. This analysis provides an overview of the sub-examiners' perspectives and levels of agreement regarding the different aspects investigated in the study.

Results

The purpose of developing and administering the questionnaire was twofold. Firstly, it aimed to investigate the strategies employed by the examiners (graders of the answer scripts) to enhance the reliability of the marking process in the assessment. The focus was on identifying the practices and approaches utilized by examiners to ensure consistent and accurate grading. Secondly, the questionnaire sought to explore any practices on the part of the examiners or the board that could potentially pose risks to the marking reliability. By gathering information on these aspects, the study aimed to contribute to the improvement of the assessment process.

The following results were obtained from item-wise questionnaire analysis.

Table 1

Academic Qualification of the Examiners

Qualification	Frequency	Percent	Cumulative percent
Graduate	16	17.4	17.4
Masters	73	79.3	96.7
M.Phil	3	3.3	100
Total	92	100	

Table 1 presents the educational qualifications of the examiners who participated in the study. 79.3 percent examiners hold an M.A. degree, indicating a significant portion of the sample has attained a master's level of education. 17.4 percent examiners are graduates, suggesting a lower but still notable percentage of examiners with a bachelor's degree. A smaller proportion, 3.3 percent, have an M.Phil degree. These findings suggest that the majority of the evaluators involved in the study are well-qualified and hold appropriate educational credentials to perform marking tasks efficiently.

Corresponding Author's Email: batool.atta@ajku.edu.pk

Table 2

Professional Qualification of the Examiners

Qualification	Frequency	Percentage	Cumulative percentage
TEFL	3	3.3	3.3
M.Ed	29	31.4	35.7
B.Ed	60	65.2	100
Total	92	100	

Table 2 indicates 3.3 percent examiners hold a Diploma in Teaching English as Foreign Language (TEFL), 31.4 percent hold an M.Ed. degree, and 65.2 percent have a bachelor's degree in education (B.Ed.). Thus, all participant examiners have appropriate professional qualifications.

Table 3

Teaching Experience of the Examiners

Experience	Frequency	Percentage	Cumulative percentage
<5 years	0	0	0
5-10 years	4	4.3	4.3
10-15 years	4	4.3	8.6
15-20 years	12	13.0	21.6
20-25 years	0	0	21.6
25-30 years	20	21.7	43.3
>30	52	65.5	100
Total	92	100	

Table 3 shows 4.3 percent examiners have 5 to 10 years of teaching experience, another 4.3 percent have 10-15 years of experience, 13 percent have 15-20 years of experience, while 21.7 percent have 25-30 years of experience, and 65.5 percent have more than 30 years teaching experience. Thus, all examiners have more than five years' teaching experience.

Table 4

Experience in Teaching English at Secondary Level

Experience	Frequency	Percentage	Cumulative percentage
<5 years	16	17.4	17.4
5-10 years	16	17.4	34.8
10-15 years	16	17.4	52.2
15-20 years	12	13.0	65.2
20-25 years	4	4.3	69.5
25-30 years	4	4.3	73.8
>30	24	26.1	100
Total	92	100	

Table 4 indicates 17.4 percent examiners have less than five years of experience in teaching English at the secondary level. 17.4 percent have 5-10 years of experience, 17.4 percent have 10-15 years' experience. 13 percent have 15-20 years' experience, 4.3 percent have 20-25 years, and another 4.3 percent have 25-30 years of experience. The most significant percentage of examiners (26.1 percent) have more than 30 years of experience in teaching English at the secondary level. Thus most participant evaluators have extensive experience in teaching English at the secondary level.

Table 5

Training taken in marking papers

Experience	Frequency	Percentage	Cumulative percentage
Yes	0	0	0
No	92	100	100
Total	92	100	

Table 5 reveals that none of the examiners in the study have received any training paper marking. This clearly suggests a lack of formal training for paper evaluation at the secondary level in AJKBISE.

Table 6

Paper marking experience

Experience	Frequency	Percentage	Cumulative percentage
<5 years	28	30.5	30.5
5-10 years	16	17.4	47.9
10-15 years	16	17.4	65.3
15-20 years	8	8.7	74.0
20-25 years	8	8.7	82.7
25-30 years	8	8.7	91.4
>30	8	8.7	100
Total	92	100	

Table 6 indicates 30.5 percent examiners have less than five years of experience in paper marking, 17.4 percent have 5-10 years of experience, another 17.4 percent have 10-15 years of experience, while 34.8 percent of examiners have a range of 15-30 and more years' experience in papers evaluation.

Table 7

English paper marking experience

Experience	Frequency	Percentage	Cumulative percentage
<5 years	32	34.7	34.7
5-10 years	24	26.1	60.8
10-15 years	8	8.7	69.5
15-20 years	4	4.3	73.8
20-25 years	12	13.0	86.8
25-30 years	8	8.7	95.5
>30	4	4.3	100
Total	92	100	

Corresponding Author's Email: batool.atta@ajku.edu.pk

Table 7 shows 34.7 percent examiners have less than five years' experience in English papers evaluation. 26.1 percent have 5-10 years' experience, 8.7 percent of examiners have 10-15 years of experience. 4.3 percent have 15-20 years, 13 percent have 20-25 years, 8.7 percent have 25-30 years, and 4.3 percent have more than 30 years experience in marking English papers. This indicates that more examiners have extensive experience in English paper marking.

Table 8

Average number of papers you mark in a session.

The average number of papers	Frequency	Percentage	Cumulative percentage
280	4	4.3	4.3
300	52	56.5	60.9
350	4	4.3	65.2
400	8	8.7	73.9
550	4	4.3	78.3
600	20	21.7	100
Total	92	100	

Table 8 shows the average number of papers that various evaluators mark in one session. Majority of the examiners mark an average of 300 papers per session. Some of them mark up to 600 papers. A bundle of papers typically contains 250 to 330 papers. It means that some examiners mark more than one bundle.

Table 9

BISE Mirpur sends a bundle of papers in your name

Response	Frequency	Percentage	Cumulative percentage
Always	32	34.8	34.8
Mostly	20	21.7	56.5
Sometimes	12	13.0	69.5
Rarely	8	8.7	78.2
Never	20	21.7	100
Total	92	100	

Table 9 indicates 34.8 percent examiners regularly receive a bundle of answer papers for evaluation from the board, 21.7 percent mostly receive, 13 percent sometimes receive, 8.7 percent rarely receive, while 21.7 percent never receive papers from the board, but they still do paper marking. It means they recieve papers from other examiners, who are either unwilling to do evaluation or the head examiners do not want them to allocate the job to them. It also shows the willingness and availability of these 21.7 percent examiners to do the marking.

Table 10

If the board does not send you a bundle, you mark papers as an alternate examiner for another examiner, who is unwilling, or has no time, or the head examiner does not want to allot him papers for some reason.

Response	Frequency	Percentage	Cumulative percentage
Always	0	0	0
Mostly	24	26.1	26.1
Sometimes	16	17.4	43.5
Rarely	20	21.7	65.2
Never	32	34.8	100
Total	92	100	

Table 10 shows 26.1 percent examiners mostly work as alternate examiners. 17.4 percent sometimes work, 21.7 percent rarely mark papers as alternate examiners, while 34.8 percent of examiners never mark papers as alternate examiners. Cross-examination of tables 17 and 18 verifies the results of both tables.

Table 11

You are completely satisfied that the answer key for objective questions provided by the board is 100 percent correct, and you exactly follow the answer key.

Response	Frequency	Percentage	Cumulative percentage
Always	36	39.1	39.1
Mostly	32	34.8	73.9
Sometimes	12	13.0	87.0
Rarely	12	13.0	100
Never	0	0	100
Total	92	100	

Table 11 shows 39.1 percent examiners are always satisfied with the key provided by the board for objective questions. 34.8 percent are mostly satisfied, 13.0 percent examiners are partially satisfied, and another 13.0 percent are rarely satisfied with the key as they may find errors. The result of this table raises questions on the authenticity of the answer key provided by the board for objective questions included with the question paper.

Corresponding Author's Email: batool.atta@ajku.edu.pk

Table 12

You are completely satisfied and agree with the guidelines for scoring constructed response test items (subjective type questions) provided by the board.

Response	Frequency	Percentage	Cumulative percentage
Always	60	65.2	65.2
Mostly	24	26.1	91.3
Sometimes	0	0	91.3
Rarely	8	8.7	100
Never	0	0	100
Total	92	100	

Table 12 indicates 65.4 percent examiners are always satisfied with the guidelines from the board for marking subjective questions. 26.1 percent are mostly satisfied, 8.7 percent of examiners are rarely satisfied, as they may find flaws.

Since most examiners follow the guidelines as these are, marking reliability must be affected if there are mistakes in the answer key/guidelines provided by the board.

Table 13

You finish scoring one complete paper at a time and then start scoring another.

Response	Frequency	Percentage	Cumulative percentage
Always	76	82.6	82.6
Mostly	8	8.7	91.3
Sometimes	4	4.3	95.7
Rarely	0	0	95.7
Never	4	4.3	100
Total	92	100	

Table 13 shows 82.6 percent examiners always finish scoring all questions of one paper before starting another. 8.7 percent of examiners mostly do the same, while 4.3 percent do this occasionally and 4.3 percent never mark all questions of a single paper.

Table 14

You score one question of all papers at a time (e.g., Question 2), then score another question of all papers (e.g., Question 3), and so on.

Response	Frequency	Percentage	Cumulative percentage
Always	4	4.3	4.3
Mostly	1	1.0	5.3
Sometimes	4	4.3	9.7
Rarely	8	8.7	18.4
Never	75	81.5	100
Total	92	100	

Table 14 shows that only 4.3 percent examiners always mark the same question in all the papers at a time, 1 percent do mostly, 4.3 percent do occasionally and 8.7 percent do rarely; while 81.5 percent never follow this technique. Table 13 and Table 14 endorse each other results.

Table 15

For long questions like summary, a body of the letter/application, story, dialogue, or paragraph/essay, you prepare rubrics first to allocate separate grades for different aspects of writing (spelling, grammar, appropriateness of vocabulary, appropriateness of length, appropriateness of language used, organization of the text, etc.).

Response	Frequency	Percentage	Cumulative percentage
Always	60	65.2	65.2
Mostly	12	13.0	78.3
Sometimes	12	13.0	91.3
Rarely	0	0	91.3
Never	8	8.7	100
Total	92	100	

Table 15 shows 65.2 percent examiners always develop rubrics to allocate separate grades for different writing and language skills before marking essay-type questions. 13 percent do mostly, another 13 percent do sometimes, and 8.7 percent never develop rubrics for detailed questions.

Table 16

While scoring papers, you remain very strict.

Response	Frequency	Percentage	Cumulative percentage
Always	12	13.0	13.0
Mostly	16	17.4	30.4
Sometimes	16	17.4	47.4
Rarely	12	13.0	60.4
Never	36	39.1	100
Total	92	100	

Table 16 indicates 13 percent examiners state that they do strict marking; 17.4 percent declare themselves mostly strict; 17.4 percent are occasionally, and 13 percent of examiners are rarely strict. 39.1 percent said they are not never strict while marking papers.

Table 17

Corresponding Author's Email: batool.atta@ajku.edu.pk

You have a soft corner for the candidates while scoring papers.

Response	Frequency	Percentage	Cumulative percentage
Always	40	43.5	43.5
Mostly	16	17.4	60.9
Sometimes	16	17.4	78.3
Rarely	0	0	78.3
Never	20	21.7	100
Total	92	100	

Table 17 shows 43.5 percent examiners are always lenient in marking, 17.4 percent are mostly lenient, and 17.4 percent are occasionally lenient. 32.7 percent state they are never lenient while marking papers.

Table 18

If a candidate reaches very close to passing, you review the answer sheet to find some points where he/she could be granted more grades to pass.

Response	Frequency	Percentage	Cumulative percentage
Always	56	60.9	60.9
Mostly	32	34.8	95.7
Sometimes	4	4.3	100
Rarely	0	0	100
Never	0	0	100
Total	92	100	

Table 18 shows if a candidate scores close to passing marks, 60.9 percent of examiners always review the answer script to accommodate the candidate to pass. 34.8 percent of examiners mostly make such an effort, while 4.3 percent of examiners occasionally accommodate such cases.

Table 19

You are extra careful in counting, writing grades in figures and words, and making an award list.

Response	Frequency	Percentage	Cumulative percentage
Always	64	69.6	69.6
Mostly	24	26.1	95.7
Sometimes	4	4.3	100
Rarely	0	0	100
Never	0	0	100
Total	92	100	

Table 19 shows 69.6 percent examiners are always extra careful in tabulation of results to avoid any mistakes at all stages. 26.1 percent of examiners mostly careful, 4.3 percent of examiners are sometimes. These results indicate that most examiners are careful in result tabulation.

Table 20

You help other examiners with their paper marking tasks.

Response	Frequency	Percentage	Cumulative percentage
Always	4	4.3	4.3
Mostly	20	21.7	26.1
Sometimes	32	34.8	60.9
Rarely	16	17.4	78.3
Never	20	21.7	100
Total	92	100	

Table 20 indicates 4.35 examiners always help other examiners with their paper marking tasks. 21.7 percent examiners mostly help, 34.8 percent admitted they occasionally help others. 17.4 percent rarely help, while 21.7 percent never help other examiners.

Table 21

The head examiner forces you to score extra papers other than those allotted to you, for which you will be paid nothing.

Response	Frequency	Percentage	Cumulative percentage
Always	12	13.0	13.0
Mostly	8	8.7	21.7
Sometimes	12	13.0	34.7
Rarely	12	13.0	47.7
Never	49	52.2	100
Total	92	100	

Table 21 indicates 13 percent of examiners reported that their head examiners always forced them to mark additional papers for them without any remuneration. 8.7 percent of examiners reported a lesser frequency, and 13 percent of examiners state occasional forced and unpaid marking. 13 percent are rarely forced to do this extra work, while 52.2 percent never forced for this work.

Table 22

Paper marking increases your workload.

Response	Frequency	Percentage	Cumulative percentage
Always	20	21.7	21.7
Mostly	20	21.7	43.5
Sometimes	32	34.8	78.3
Rarely	12	13.0	91.3
Never	8	8.7	100
Total	92	100	

Table 22 shows 21.7 percent examiners report paper marking always increases their workload. Another 21.7 percent examiners mostly found increase

Corresponding Author's Email: batool.atta@ajku.edu.pk

in their workload, 34.8 percent examiners state it sometimes, and 13 percent say that this task rarely increases their workload, while 8.7 percent state that marking never increases their workload.

Table 23

You get tired/feel burdened in the days of paper marking.

Response	Frequency	Percentage	Cumulative percentage
Always	8	8.7	8.7
Mostly	28	30.4	39.1
Sometimes	32	34.8	73.9
Rarely	16	17.4	91.3
Never	8	8.7	100
Total	92	100	

Table 23 shows 8.7 percent examiners always get tired or feel burdened during paper marking. 30.4 percent examiners mostly, 34.8 percent sometimes, 17.4 percent rarely, and 8.7 percent never get tired or feel burdened in the days when they mark papers.

Table 24

You wish someone could help you with paper marking.

Response	Frequency	Percentage	Cumulative percentage
Always	12	13.0	13.0
Mostly	8	8.7	21.7
Sometimes	4	4.3	26.0
Rarely	20	21.7	47.7
Never	48	52.2	100
Total	92	100	

Table 24 indicates 13 percent examiners always wish for help for paper marking, 8.7 percent mostly, 4.3 percent sometimes, and 21.7 percent rarely wish for assistance in paper marking. The majority examiners (52.2 percent) never feel the need of assistance.

Table 25

You can find someone in your family/friends who can help you.

Response	Frequency	Percentage	Cumulative percentage
Always	12	13,0	13.0
Mostly	0	0	13.0
Sometimes	8	8.7	21.7
Rarely	12	13.0	34.7
Never	60	65.2	100
Total	92	100	

Table 25 indicates 13 percent examiners always get assistance from their family or friends in paper marking. 8.7 percent sometimes, thirteen percent rarely get assistance, while 65.2 percent never get assistance in paper marking.

Table 26

The person(s) among your friends/family, who help(s) you in scoring papers, is equally qualified/experienced as you are.

Response	Frequency	Percentage	Cumulative percentage
Always	3	3.3	3.3
Mostly	17	18.5	21.9
Sometimes	0	0	21.9
Rarely	11	12.0	33.9
Never	60	65.2	100
Total	92	100	

Table 26 indicates 3.3 percent examiners state that the persons who assist them in marking always equally qualified and experienced as themselves, 18.5 percent state that most helpers are equally qualified. 12 percent state that their helpers are rarely equally qualified/experienced, and 65.2 percent examiners reported 'never' in their questionnaire.

Cross-examination with table 25 verifies that 65.2 percent of examiners do not seek assistance for marking.

Table 27

The person(s) among your friends/family, who help(s) you in scoring papers, makes mistakes in scoring.

Response	Frequency	Percentage	Cumulative percentage
Always	0	0	0
Mostly	16	17.4	17.4
Sometimes	8	8.7	26.1
Rarely	8	8.7	34.8
Never	60	65.2	100
Total	92	100	

Table 27 indicates 17.4 percent examiners admit that their assitants mostly make mistakes in marking, 8.7 percent state that helpers sometimes make mistakes. Another 8.7 percent examiners report that their helpers rarely make mistakes. Since 65.2 percent of examiners do not seek help from anyone in the marking job, they selected 'Never.'

Table 25, 26 and 27 also verify the consistency in responses as reported by the participants.

Corresponding Author's Email: batool.atta@ajku.edu.pk

Table 28

The head examiner is fully satisfied with your scoring job.

Response	Frequency	Percentage	Cumulative percentage
Always	56	60.9	60.9
Mostly	32	34.8	95.7
Sometimes	0	0	95.7
Rarely	4	4.3	100
Never	0	0	100
Total	92	100	

Table 28 indicates 60.9 percent examiners claim complete satisfaction of their head examiners with their (examiners') job, 34.8 percent say that their head examiners are mostly satisfied with their marking, 4.3 percent admit that their heads are rarely satisfied with their marking.

Table 29

The head examiner sends some papers back to you, giving you instructions to correct the mistakes pointed out.

Response	Frequency	Percentage	Cumulative percentage
Always	0	0	0
Mostly	8	8.7	8.7
Sometimes	20	21.7	30.4
Rarely	40	43.5	73.9
Never	24	26.1	100
Total	92	100	

Table 29 indicates 8.7 percent examiners state that their head examiners mostly send some papers back to them for correction. 21.7 percent of examiners say that their head examiners sometimes send them back, 43.5 percent claim that their heads examiners rarely send papers back for corrections. 26.1 percent examiners never received any papers never for correction.

Table 30

You are satisfied with the amount of remuneration paid to you by the board for scoring papers.

Response	Frequency	Percentage	Cumulative percentage
Always	12	13.0	13.0
Mostly	20	21.7	34.7
Sometimes	8	8.7	43.4
Rarely	4	4.3	47.7
Never	48	52.2	100
Total	92	100	

Table 30 indicates 13 percent examiners are always satisfied with the remuneration paid to them for paper marking, 21.7 percent state they are mostly satisfied, 8.7 percent are sometimes satisfied, and 4.3 percent are rarely satisfied. In contrast, the majority of examiners, i.e., 52.2 percent, are never satisfied with the remuneration paid to them for paper marking.

Table 31

The time you get paid by the board after scoring papers is reasonable.

Response	Frequency	Percentage	Cumulative percentage
Always	8	8.7	8.7
Mostly	0	0	8.7
Sometimes	4	4.3	13.0
Rarely	0	0	13.0
Never	80	87.0	100
Total	92	100	

Table 31 indicates 8.7 percent examiners are always satisfied with the timeline they get paid by the board after marking papers, 4.3 percent examiners find the timeline sometimes reasonable. A large majority of the examiners, i.e., 87 percent, state that the timeline of getting paid by the board is never reasonable. It suggests that the board delays evaluators' remunerations.

Table 32

You score the papers to compensate for your financial deficiencies.

Response	Frequency	Percentage	Cumulative percentage
Always	17	18.5	18.5
Mostly	45	48.9	67.4
Sometimes	3	3.3	70.7
Rarely	7	7.6	78.3
Never	20	21.7	100
Total	92	100	

Table 32 shows 18.5 percent examiners always take evaluation work to supplement their income, 48.9 percent also admit that they mostly do marking to supplement their income, 3.3 percent do this sometimes for financial reasons. 7.6 percent examiners say they rarely mark papers for financial reasons, 21.7 percent examiners however claim that they never do marking to supplement their income.

Corresponding Author's Email: batool.atta@ajku.edu.pk

Table 33

You score the papers because you think you can do this well, want to contribute to educational services efficiently, and enjoy it.

Response	Frequency	Percentage	Cumulative percentage
Always	55	59.8	59.8
Mostly	16	17.4	77.2
Sometimes	17	18.5	95.7
Rarely	0	0	95.7
Never	4	4.3	100
Total	92	100	

Table 33 indicates 59.8 percent examiners claim that they always mark papers for the reason that they think that they can do that well, to contribute to educational services, and/or they enjoy it, 17.4 percent of examiners say that they mostly do the marking job for the reasons mentioned above, 4.3 percent examiners admit that they never mark the papers for these reasons.

Discussion

The purpose of developing and administering the questionnaire was to investigate the strategies incorporated by the examiners (graders of the answer scripts) to improve the grading reliability of the assessment process, as well as to investigate any practices on the part of examiners or the board, that may be hazardous to marking reliability. The following implications emerged from the analysis of questionnaire data.

The results revealed that almost all the examiners are academically and professionally qualified (Tables 1 & 2). The majority of examiners also bear long teaching experience (Table 3), including experience in English teaching (Table 4). It is important to note that no training/workshop is done by the board or education department to train the examiners in paper marking (Table 5). Training in paper marking may develop uniformity in the award of grades by different examiners. The absence of such training is a challenge to marking reliability because it may lead to differences in awards or grades by different evaluators. Results also revealed that the majority of the examiners are well-experienced not only in general paper marking (Table 4) but also in marking English papers (Table 6).

The average number of papers allotted to the majority of examiners for marking is 300, although some examiner mark as many papers as 600 (Table 8). It seems very challenging for one person to mark so many papers. There are two risks in marking a large number of papers; one is that the examiner may give some papers to other people near him for marking, who may not be able to mark papers reliably; secondly, if the examiners mark these papers themselves, it must increase their workload to an unbearable level, and this also may be a risk to reliability.

BISE Mirpur does not always send papers for marking to all the examiners (Table 9), so a considerable number of examiners mark papers as alternate examiners (Table 10). This shows that many examiners are eager to do the marking job.

The majority of the examiners are satisfied with the answer key (Table 11) and guidelines (Table 12) provided by the board. Some examiners are rarely or only sometimes satisfied with the key and guidelines. It implies that there are sometimes mistakes in the key, and some examiners spot those mistakes. On the other hand, most examiners are completely satisfied with the key and guidelines, which implies that they cannot spot the mistakes their fellow examiners find. The existence of mistakes in the key, but most examiners cannot spot them, makes the marking process less reliable.

Most examiners finish marking all questions on a single paper at a time (Table 13). Few examiners simultaneously mark the same questions on all papers (Table 14). A hazard for marking reliability emerges from the cumulative result of these two tables. The marking of one question in all papers is recommended by experts (Gipps, 2002), as it improves

Corresponding Author's Email: batool.atta@ajku.edu.pk

the marking reliability, but it is lacking in the current situation.

The majority of the examiners prepare a rubric or scoring scheme before marking long questions (Table 15), which is a desired practice, and improves the reliability of subjective questions. Regarding strictness or softness in marking papers, results show moderate behavior of the examiners. There is a scattered result, as shown in Tables 16 and 17. This situation improves in the case of a candidate who reaches near-passing marks but requires a few more to pass. In such cases, most examiners review the answer paper to find some question/point, where the candidate may be given a few more marks so that he/she could pass. Uniformity in examiners' behavior is seen in this regard (Table 18).

Most examiners are extra careful in counting grades, writing them in figures and words, and making award lists (Table 19). There are a few examiners who selected 'mostly' or 'sometimes.' This means that chances of mistakes in the compilation of results also exist. Workload may also be a challenge to the reliability of marking. Due to workload, examiners seek help from others in marking jobs. Persons who help in marking are sometimes less qualified, less experienced, and inefficient. Thus, the reliability of marking suffers.

A considerable percentage of the examiners admit that they help other examiners in their paper marking tasks (Table 20). BISE often allots a bundle of papers to head examiners themselves for marking. Some head examiners distribute the papers to sub-examiners to mark those for them. Examiners are not paid for such extra marking tasks. This practice also increases their workload and fatigue (Table 21).

Paper marking is a tough job, and increases the workload when done with a regular job. This increased workload may not only make the examiners tired but also may affect their marking efficiency, and their marking may be less reliable in such circumstances. Many examiners expressed that marking jobs increase their workload (Table 22). Many of the examiners feel tired on the days of paper marking (Table 23). Few wish someone in their family or friends could help them with this task (Table 24). Some of them even take the help of others in their marking task (Table 25). These helpers are mostly less qualified and less experienced, so they make mistakes in marking (Table 26, 27). These facts lead to a loss of marking reliability.

Most of the examiners are efficient enough, and the head examiners are always/mostly satisfied with their marking job (Table 28), but some of them are not much efficient, and the heads are not always satisfied with their work. They sometimes make mistakes in marking, and heads send such papers back to them for corrections (Table 29). So, the inefficiency of the examiners is also a risk to marking reliability.

When a person gets paid for a task, he/she performs that task with great interest. The more one is paid, the better he/she will do the job. BISE also pays remuneration to the examiners for paper marking after about seven to eight months, but sometimes this duration is even more. Most of the examiners are never satisfied with the remuneration paid to them (Table 30). They are also not satisfied with the time after which they receive the remuneration (Table 31). The majority of the examiners do the marking job to increase their income (Table 32), although the majority also claim that they do this job because they want to serve the education and enjoy the job (Table 33). This claim contradicts one of the previous statements, which says that most examiners get tired and feel burdened while marking papers. It is more probable that examiners undertake and perform the marking task to enhance their income, not to serve education. When they are paid less or paid late, they may lose interest in

Corresponding Author's Email: batool.atta@ajku.edu.pk

marking, and the reliability of marking may suffer.

To sum up, good things about the grading of answer scripts of secondary school level English in Azad Jammu and Kashmir are that the examiners are well qualified academically and professionally, they are well experienced not only in teaching but also in grading, they keep a soft corner for the candidates while grading, they make a rubrics/scoring scheme for subjective questions and they are careful in counting, etc. Unwanted and problematic practices in the marking process are that examiners are extra eager to get papers for grading, they grade whole papers at a time and do not mark question-wise, grading task makes them tired and burdened, helping others (especially the head examiners) makes them even more tired and bored, they are unsatisfied with the rate of remuneration and the time in which it is paid, and they sometimes get the help of less qualified and inefficient persons. There are also chances of mistakes in answer keys and guidelines the board provides, and most examiners follow these blindly.

Conclusion

Although the examiners who grade the answer scripts are well-qualified and experienced, no training is given to them in grading answer scripts. Some of them are less responsible, and the increased workload leads them to put reliability at risk. The board may provide rubrics for the evaluation of subjective questions. This is necessary to ensure uniformity of awarding grades to all candidates and to improve inter-rater reliability of assessment. Moreover, such a practice may train new and un-experienced examiners. It may also make the students focus on all aspects of academic writing. Training/workshops for examiners are strongly recommended; papers may only be allotted to trained examiners. In such workshops, the examiners should be provided guidelines for uniform grading of answer scripts. They may be trained to make their rubric for subjective test items if not provided by the board. These workshops may discuss significant issues, and valuable suggestions may be obtained from the trainees and trainers.

References

Abbara, T. M. (2004) Testing English as a foreign language: a case study of classroom tests in Qatar. Qatar: Durham University.

Baartman, L. K., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. (2007). Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational Research Review*, *2*(2), 114-129.

Brennan, R. L. (2006). *Educational Measurement. ACE/Praeger Series on Higher Education*. Praeger. Available from: Greenwood Publishing Group, Inc. P.O. Box 5926, Portsmouth, NH 03802-6926.

Carr, N. T. (2011). *Designing and analyzing language tests: Oxford handbooks for language teachers*. Oxford, U.K.: Oxford University Press.

Dorans, N. J., & Cook, L. L. (Eds.). (2016). *Fairness in educational assessment and measurement*. Routledge.

Fives, H. & Barnes, N. D. (2013) *Practical Assessment, Research & Evaluation* Volume 18, Number 3, February 2013

Gebril, A. (2016). EDUCATIONAL ASSESSMENT IN MUSLIM COUNTRIES. *Handbook of human and social conditions in assessment*, 420.

Ghazali, N. H. M. (2016). A Reliability and Validity of an Instrument to Evaluate the School-Based Assessment System: A Pilot Study. *International Journal of*

Corresponding Author's Email: batool.atta@ajku.edu.pk

Evaluation and Research in Education, *5*(2), 148-157.

Gipps, C. (2002). *Beyond testing: Towards a theory of educational assessment*. Routledge.

Gitomer, D. H., Martínez, J. F., Battey, D., & Hyland, N. E. (2019). Assessing the assessment: Evidence of reliability and validity in the edTPA. *American Educational Research Journal*, 0002831219890608.

Joseph, G., Soderberg, J. S., Stull, S., Cummings, K., McCutchen, D., & Han, R. J. (2020). Inter-rater reliability of Washington State's kindergarten entry assessment. *Early Education and Development*, *31*(5), 764-777.

Kuramoto, N., & Koizumi, R. (2018). Current issues in large-scale educational assessment in Japan: Focus on the national assessment of academic ability and university entrance examinations. *Assessment in education: principles, policy & practice*, *25*(4), 415-433.

Kyani, A. (2011) An Analysis of Scoring Secondary Level Mathematics for Board of Intermediate and Secondary Education Sargodha, and its Impact on Reliability of the Examination. Islamabad: AIOU

Lockwood, A. B., Sealander, K., Gross, T. J., & Lanterman, C. (2020). Teacher Trainees' Administration and Scoring Errors on the Kaufman Test of Educational Achievement. *Journal of Psychoeducational Assessment*, *38*(5), 551-563.

Malone, M. E. (2017). Training in language assessment. *Language testing and assessment*, 225-239.

Murphy, R. (2006). Evaluating new priorities for assessment in higher education. *Innovative assessment in higher education* (pp. 57-67). Routledge.

Phelan, C. & Wren, J. (2006) Exploring Reliability in Academic Assessment. Graduate Assessment: UNI Office of Academic Assessment. https://chfasoa.uni.edu/reliabilityandvalidity.htm

Shepard, L. A. (2009). Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practice*, *28*(3), 32-37.

Teachers Guide to Assessment (2014). Australia; ACT.

Thissen, D., & Wainer, H. (2001). Test Scoring. Mahwah, NJ: Erlbaum Trevor Michael Edward Forde, Edith Cowan University, 1993.

Timothy, M. (2007) Construction, validation, and administration of a diagnostic test for undergraduates. Florida: University of Florida.

William, D. (2013) Assessment: The Bridge between Teaching and Learning. *Voices from the Middle*, Volume 21 Number 2, December 2013.

Corresponding Author's Email: batool.atta@ajku.edu.pk