How much reliable are our language tests? A case study

Mohammed Humayun Kabir*

Abstract: In this paper we will first attempt to familiarize the paradigm shifting in language teaching in Bangladesh. Then, the umbrella term reliability is discussed with references to the literature. Subsequently, four different data instruments have been analysed in order to investigate how far reliability is maintained in testing Reading and Writing by taking HSC (Higher Secondary Certificate) English testing as sample. The study found that reliability at HSC English Testing is not maintained, as item setting, marking, test administration etc. are problematic. Finally, this study recommends some pragmatic measures like providing rater's training, introducing rating scale, establishing fair and impartial test administration, etc. to achieve the desired goal.

Introduction

In Bangladesh our Ministry of Education introduced Communicative Language Teaching in Secondary School Certificate (SSC) and Higher Secondary Certificate (HSC) curriculum in 2000-2001. But resistance to it, particularly from a section of teachers, was immediate. 'When CLT came to Bangladesh the traditional English teachers vehemently opposed it because they were not ready for something new', Selim and Mahboob, (2001:141). In fact this sort of teacher resistance is not unusual. While evaluating Pennington's model Canagarajah (2002:137) also predicts that '..there could be significant teacher resistance to new methods and that the values/interests/predispositions of the teachers will mediate the reception of the new method.' Experiments after experiments are going on in Bangladesh in order to impart effective English language teaching (Chapter-1 of the dissertation done by Islam: 2003, University of Essex, UK). Hoque (2002), ELT advisor to Bangladesh Open University forecasts that 'most teachers not trained in CLT would find it difficult to teach and test their students'. After observing the present situation, Shahidullah (2003) points out some factors which are the main obstacles for the effective language learning and teaching where he identifies 'wrong testing methods' as one of those factors. 'Our examination system (setting of questions, marking papers, learning outcomes, etc) needs to be modified ....' Rozario (2005). The test users can hardly rely on the test results and most of the time they are affected by it. All these features are certainly *reliability* worry. Davies (1990:1) emphasizes that

* Mohammed Humayun Kabir, Assistant Professor, Department of English Language and Literature, International Islamic University Chittagong (IIUC)

'language testing is central to language teaching.' Schwartz (2002:128) states 'many attempts at curricular change have been foiled due to the lack of corresponding change in method of assessment' as cited in Islam (2003).

As I am a language teacher and tester I have a great curiosity to examine the reliability of our language tests. That is why here this research will deal with Higher Secondary Certificate (HSC) English testing in Bangladesh as a sample.

I expect this study will identify the prevailing *reliability problem* in our testing at HSC and it will give an overall idea about the language tests in Bangladesh at school and college levels.

1.2. Structure and Marks distribution:

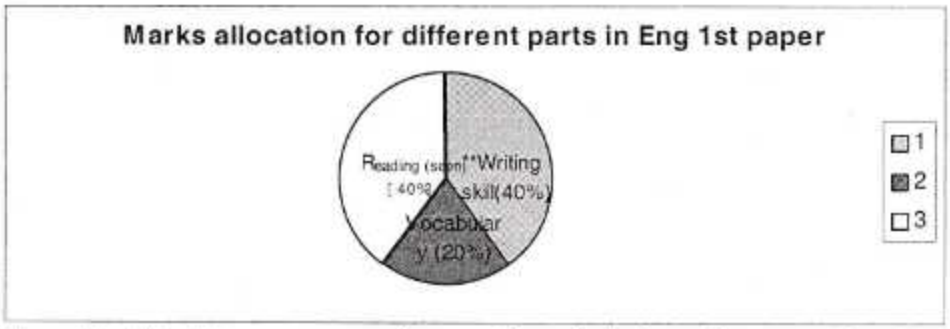

Figure 1.1: [**writing components : Re-arranging, substitution table, paragraph writing.]

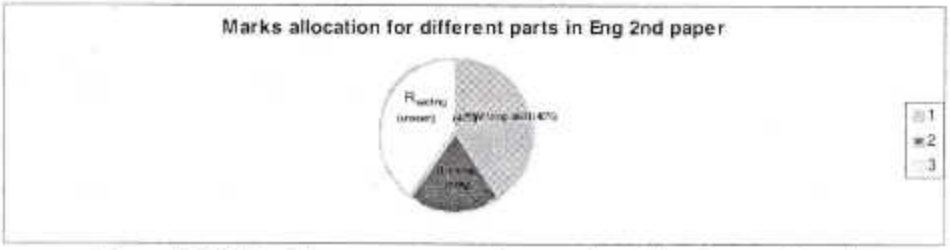

Figure:1.2, [**writing components : Paragraph writing, letter writing / writing creatively from experience, completing story / continuing passage]

1.3. Scoring:

At HSC, both subjective and objective markings are done. When the subjective marking is done, raters are advised to take into account the following criteria during assessment: 'a) task fulfilment ... b) The mechanics of good writing ..'. (H.S.C Teachers' Guide).

It should be noted that there is no record of separate marking or grading for Reading / Writing skills in the final *Academic Transcript* of the student. Students are awarded one of the following grades: A+, A, A-, B, C, D, or F, which is a

sum total of the numerals s/he scores on different questions. So we see that global rating scale is followed here.

2.1.Reliability:

A test is reliable when it measures consistently. 'The reliability of a test concerns the accuracy and trustworthiness of its results,' Allison (1999:85). It is defined as the consistency of measurement. Bachman and Palmer (1996:21-23) state that reliability is clearly an important quality of test scores. If a test's scores are not consistent they cannot provide us with any information at all about the ability of a test taker we want to assess. They also opine that it (reliability) is a necessary condition for construct validity. Hughes (2003) identifies two main sources of inaccuracy in language ability testing, which are 'test content and test techniques and lack of reliability'. It is obvious that 'the better all of the stakeholders in test or testing system understand testing, the better the testing will be' (ibid:5).

Hughes (2003:44) also has set two components of test reliability: i) the performance of candidates from occasion to occasion and ii) the reliability of the scoring.

The HSC test which I have taken into consideration in my research is a massive one. This year (2009) about 6,20,000 students are taking part in the HSC examination (sources: *The Daily Star*, the daily *Prothom Alo* of 16th April) under seven education boards across the country. In such a large test we can look for the second component that Hughes (2003) has mentioned. So here I will confine my discussion only to the reliability of the marking.

Hughes (2003), Weir (2003), Weir (2005), Allison (1999), Anastasi (1988) and Alderson et al (1995), Alderson, J.C (2000), White (1984), have recommended some valuable means of achieving reliable performance from test takers:

i) Items should be set that permit scoring which is as objective as possible.

ii) A detailed scoring guideline should be provided.

iii) Scorers' training is a must.

iv) Multiple and independent scoring should be employed.

v) Intra-rating reliability i.e. *'consistent within himself'* or herself is a very crucial issue of reliability where a single marking is considered as final marking.

vi) Examiners need to become familiar with the marking systems (schemes or scales) that they are expected to use and they must learn how to apply them consistently.

vii) Candidates should be made familiar with format and testing techniques.

In this study I am going to investigate most of the above-mentioned points.

2.2. Training of the markers and test administrators:

In Bangladesh there is no formal training for the markers. Usually as soon as the exam (HSC) finishes all examiners are called in the education board office and there is a briefing session for the examiners which is conducted by the Controller of the Examinations and by the Head Examiners where the examiners are urged to mark carefully. Then a 'key' is handed to all examiners. In this way the session comes to an end.

In fact, it is the examiners who can make or mar the reliability of a test to a large extent.

Apart from the examiners, test administrators also need training. Though this training is not as complicated and lengthy as the training of the raters, 'it is still important that the administrators understand the nature of the test they will be conducting, the importance of their own role and the possible consequences for candidates if the administration is not carried out correctly' (Alderson et al, 1995:115). To make sure fair, consistent and reliable marking, we also need to monitor the marking done by the markers.

2.3. Monitoring examiner reliability:

Scores on a test should be rater-independent. If marking is rater-dependent i.e. marking varies significantly from time to time (within a marker) or person to person then that marking is unreliable. 'An unreliable examiner is somebody who changes his or her standards during marking, who applies criteria inconsistently , or who does not agree with other examiners' marks', Alderson et al (1995:128). If we look for the answer why inconsistent and unreliable judgment occur we will find several factors: (a) problem with rating scale (b) time pressure (c) domestic and professional worries, etc. It is imperative to monitor the marking of a test in order to increase the reliability of the marking. Sampling the examiner's marks and asking for judgements to be made if the marking is not satisfactory are the most recognised ways of ensuring reliable marking. Double- marking is also another well-appreciated and widely-accepted means of reliable marking. HSC does not have dependable system of monitoring of the marking. (See questionnaire analysis below).

2.4. Teaching and Testing:

In Communicative Language Teaching, Breen and Candlin (1980:99) suggest that teachers may play three key roles. The first role 'is to facilitate the communication process between all participants in the classroom...', the second role is 'to act as an independent participant within the learning-teaching group' and the third role is 'that of researcher and learner'. Richards and Rodgers (2001:167) agree with the above-mentioned roles of teachers and add 'other roles assumed for teachers are needs analyst, counsellor, and group process manager'.

After teaching, the question of testing comes very naturally. Rea (1978) (cited in Carroll) suggests 'communicative language tests for personal and professional purposes imply a set of testing *tasks* which reflect and stimulate actual linguistic demands within specified domains'. Bachman and Palmer (1996) discuss two main purposes for language tests – the primary purpose is to make inferences about language ability, and the secondary purpose is to make decisions based on those inferences. But if there is mistrust of tests and of the testers, we can hardly rely on the test's results.

The following discussion includes the research questions which are related to reliability of testing reading and writing at HSC.

3. Research Methodology:

3.1. Research questions:

The following questions were asked:

i) To what extent is the testing of Reading and Writing at the HSC reliable?

ii) Is the Board of Education doing enough to ensure reliability of the HSC results?

iii) Are our examiners consistent enough to assess the answer scripts reliably?

3.2. Instruments:

In order to find the answers of the above mentioned questions I have used the following instruments:

a) One questionnaire addressed to the selected teachers & examiners (59) of higher secondary schools and another questionnaire addressed to the Chairman / Controller of examination of Chittagong Board of Education (CBE).

b) The official instructions to HSC English examiners.

(c) Previous test papers (Board questions & questions of different colleges).

d) Assessed scripts of different (three) colleges.

In addition to these above documents, I used some more official papers and reports to support my findings.

4. Data analysis:

4.1. Reply to the Questionnaire:

Among the 59 questionnaires distributed, I got back 25. In those responses I found that the teachers' / examiners' experience varied from 1 year to 18 years. I got reply from the teachers belonging to the colleges located in the main areas of the city to the remotest places (Rangamati, Khagrachori, West Banshkhali, etc) under Chittagong education board. In this discussion we will use T1,T2, T3…..to refer to the teacher-participants for the sake of their anonymity. I introduce

categorization phase of the questionnaire data on the basis of closed – ended questions below:

Title: Teachers' responses to questionnaire:

Questions	Yes	%	No	%
1. Teachers are following Teacher's Guide (T.G)	16	64%	9	36.%
2. Equal standard of test administration	10	40%	15	60%
3. Marking varies from examiner to examiner	22	88%	3	12%
4. Following any marking guide	14	56%	10	40%

Table – 4.1

From the above table and questionnaire data we find that 64% teachers are following T.G. While 36% teachers honestly confessed that they do not follow. Those who are using it have written that they are using teaching purpose except T4. He writes 'sometimes I need to follow T.G., because it gives the directions how to prepare students for examination'.

It is interesting to observe that when Khan (2005), an ELTIP teacher trainer and researcher points out that 'teachers are overburdened with heavy workloads with little time to spare for lesson planning, class preparation or correction of written work, access to teacher's guides is nil…..'.

In the collected responses it is revealed that 40% teachers think that there is fair and impartial administration across the country. Their comments include:

T4: 'At present all exam centres have to maintain fair administration'….. 'students cannot afford to get any unfairness'.

T9: 'all the centres are same whether it is town or village'. … 'No, students are getting any chance for doing copying'.

On the other hand, 60% respondents give a totally contradictory picture. T6, T7, T8, T11 & T19 state that 20-30%, 15%, 30%, 10% & 10% exam centres respectively fail to ensure fair and impartial test administration and students enjoy undue facilities like cheating, side talking, bringing written answers from outside and attaching them with the main script, etc. The national dailies also confirmed it.

In the questionnaire data (Q.14) it is also depicted that marking varies significantly from markers to markers. 88% participants acknowledged it. When asked why it happens they replied:

T2: 'some examiners are not competent enough'.

T4: 'difference in experience and outlook'

T7: 'most of the examiners do not know the marking scale and standard marking system'.

Others opined that it occurs due to the mentality of the teachers.

Since there is no specific marking guidelines for narrative / descriptive answers it is found that different examiners mark from different points of view (see section-4.3).

T12 & T13 feel the necessity of marker's training for the sake of reliable marking when they opined :

'training programme for the teachers again and again'....

'training programme for the teachers and the examiners should be arranged'

Responses to open-ended questions:

Q.4 asks about the specific criteria to be an examiner. 10 participants could not provide any information; they simply wrote 'I do not know'. 1 participant did not respond. 14 participants wrote their own opinions (not the education board provided criteria), though all of them are marking scripts regularly.

T4: 'Teacher should be from a recognized institution and must have experience and sincerity'. This reply seems to me self-contradictory as the information he provided in the questionnaire data reveals that he has been marking HSC scripts for 10 years and his length of teaching experience is 10 years as well. Moreover, in my study there are respondents from various institutions which even cover some very little known institutions. However, T8 provides some shocking information when she writes that apart from long experience 'good relationship with (Education) board authority' is needed to be an examiner.

Inquiry related to marking reading / writing skill (Q. 8):

For example, while marking reading:

T9: 'look for basic knowledge and originality'

T2: 'ability of writing creatively'.

Inquiry related to marking writing skill (Q. 9):

T4 : 'grammatical and language PERFECTION'

T2 : 'ability of writing answer in simple English'.

From the above data it is clear that HSC testing is in a grave situation. T.G. which is a very useful tool is not properly used by the teachers. Rater training which is a fundamental requirement is still not facilitated except the 1/2 hour orientation program in a hall room for the markers just before they receive the answer scripts. It is also clear that markers do not have any detail marking

guidelines and marking scale as well. There are no clear criteria for who can be a Board examiner.

Furthermore, the reports of national dailies (*The Daily Star, Daily Prothom Alo*) of 11-05-07 & 14-05 -07 made public that 303 students for cheating and 15 teachers for neglecting duties were expelled from exam halls on the day of English First Paper exam and 180 students and 4 teachers were expelled on the day of English Second Paper Exam for the same offences in 2007.

However, from the responses of the Controller of Examination we get totally a different idea.

When asked :

(Q. 1) Do you think all exam centres have equal standards of fair and impartial administration during examination? If not, what is the approximate percentage of those centres?

He responds: 'YES'.

(Q. 2) Do you think students get some undue facilities like copying/cheating in some exam centres? If yes, what is the approximate percentage of those centres?

He responds: 'NO'.

Q. 5. asks: Do you think marking (significantly) varies from examiner to examiner? If yes, please write why it happens.

He replies: 'I think it does not occur usually'.

It is hard to accept the responses because teachers / markers and daily newspapers' reports do not support it. We assume that as he is holding a very responsible position he cannot admit all the anomalies and irregularities that exist.

We can perceive that HSC testing is seriously lacking *reliability*. Since the number of respondents is very small, the results can only be indicative, rather than definitive.

4.2. An Official Marking Instructions to HSC English examiners:

As usual, every year after the ending of the HSC examination every marker is provided with a written document i.e. 'Instruction to Examiners' by the concerned education board authority. In this study I have taken the 'Instruction to Examiners' 2006 provided by CBE as a sample instrument to analyse.

Instruction-I states 'solutions to most of the problems have been provided in the 'Specific Instruction'. 'But if any other solutions than these offered by examinee seem to be also appropriate or suitable, then give proper value to them'. This sort of instruction might have given rise to some serious problems during marking. The less confident markers will certainly be confused by the different answers

provided by the students and the stubborn marker will not accept any possible alternative answers. So, how can the test takers rely on the marking? Above all, I also maintain that 'test writers should not rely on the students' power of telepathy to elicit the desired behaviour', Hughes (2003:47).

Instruction-m is about despatching of marks and scripts. Here examiners are asked 'to send first ten examined scripts to the respective Head Examiners within three days of the receipt of the scripts for ascertaining the proper evaluation of the scripts. The present researcher suspects that the assigned examiners might check those ten scripts with some special care as they are aware that those scripts will be scrutinized by the Head Examiner and on the basis of satisfactory marking the assigned marker will be allowed to mark the rest of the scripts. So an extra alertness is expected. Alderson et al (1995:134), also agree that 'there is always the possibility that the markers will mark the 'reliability scripts' more carefully than other scripts, and that the Team Leader (Head Examiner) will not be getting a true picture of the marker's ability to adhere to the marking scale under normal conditions ....'

In the official instructions we can find that in the second instalment examiners will mark 90 scripts, and in the third and fourth instalments 150 scripts each time. For each instalment the given time is seven days. Very reasonably we doubt how it can be possible to *mark reliably* so many scripts within such a small span of time when teachers are burdened with such heavy workload? As Quader (2001) observes, 'our teachers have a lot of problems as they are to take 25 to 30 classes per week, manage classes of 30 to 90 students at a time, and take care of administrative work also. After doing all these jobs, they have to do some private tuition to survive as their salary is very poor'. Should the authority not realize that each individual script appears with different sorts of contents? We, therefore, have logic to presume that this hurry in marking will harm the *reliability in marking*.

In the instructions it is advised that examiners should not give more than 75% marks (for Q11, Q12) if any student writes (paragraph/composition) without a title. Neither the T.G. nor the syllabus for HSC textbook has mentioned it. Why is it not well circulated? Why are only the markers informed about this? It is as if the education board authority as well as the examiners are eagerly waiting to trap the students. Here again the question of *reliability in marking* arises. Alderson et al (1995:37) also agree that 'a knowledge, for example, of the specifications written for item writers, a detailed understanding of the criteria used for marking, and familiarity with the examiners' views of students' sample answers would be invaluable for all test users and would increase the *reliability of the tests'*.

Instruction 6 for English 2nd Paper says 'if any sensible answer is found irrespective of instructions it should be awarded due credit'. On the one hand the

syllabus encourages real-life communication and linguistic competence where they discourage memorizing or cramming the answers and, on the other hand, they insist that due credit should be given for sensible answers. These answers will vary from student to student, as they will have a lot of freedom to write and 'such a procedure is likely to have a depressing effect on the *reliability* of the test', Hughes (2003:45).

The above data analysis confirmed that the marking guide is problematic as the instructions are puzzling and insufficient and it is not enough to safeguard *reliable* marking.

4.3. Previous test papers

We have taken English First Paper 2006 of Chittagong Board as a sample to analyse.

Reading skill testing:

In the question we found that both Q.1 (a, b, c, d, e), Q.2 (a, b, c, d, e) are problematic as 'students have a 50% percent chance of getting the answer right by guessing alone', Alderson (2000:222).Thus there is also a ***reliability*** problem.

Q5 (five short answer questions), Q7 (a summary writing) & Q.8 (making a flow chart), are subjective and Q6 (fill in the gaps) is objective in nature. So here again there are *reliability* concerns during marking. We can predict serious marking problem in Q8 (making flow chart), as it is very controversial in nature and it has no absolutely fixed answer. Marking Q5 (short answer questions) will also be problematic. Now we will see what the marking instruction says:

No. of the questions	Marking instructions
Q. No. 5	Full marks may be given for satisfactory and accurate answers.
Q. No.6	'Key' is given.
Q. No.7	Summary written in examinees' own language / words should be given full credit.
Q. No.8	Correct information written in short phrases within the boxes should be assigned full marks.

Figure: 4.1

It is acknowledged that 'the objectivity of scoring depends upon the completeness of the answer key.....', Alderson (2000:227). As the marking guideline does not provide 'key' for Q.5, Q.7 & Q.8, we presume serious scoring problem. Scoring the Q.7 (summary writing) may be problematic. 'Scoring the summary may, however, present problems: does the rater count the main ideas in

the summary, does she rate the quality of the summary on some scale?', Alderson (2000:232), further states 'agreeing on the main points in a text may prove well nigh impossible, even for 'expert readers' while students produce a summary'. We will find it confirmed in the analysis of the marked answer scripts later (see 4.4 below). Marking the answers of the Q.5, Q.7 & Q.8 is beset with the problems as 'Instructions to Examiners' do not give any specific instructions. So *reliable marking* might not be possible here.

Writing test:

In Q.12. (**re-arranging**), students are to re-order or re-arrange 14 sentences where they have to maintain sequence.

Q.13. Paragraph writing.

There is no marking guideline to mark Q13 in the marking instructions. It is felt that examiners need some marking guidelines to assess this answer.

To turn now to the HSC Board questions (of Chittagong 2006) of English 2nd Paper we have found the following items in reading and writing skill testing:

Reading test:

It is almost similar to the English 1st Paper in terms of item selection. In the 'key' for Q.5, it says 'answer to each question should be in one sentence, if written in more than one sentence, it may be given 50%marks'. What does 'one sentence' mean? Its length? How will the markers follow it? So we assume that *unreliable* marking may occur here too.

Writing test:

All education boards have uniformity in testing writing skill. They set the following items: paragraph writing, continuing a passage/completing a story, letter writing, writing creatively from experience. Since there is a discussion on the reliability of the testing below, I am not discussing those here.

All of these items are marked on the basis of subjective marking. We assume subjective marking without clear guidance also paves the way for *unreliable assessment.*

4.4. Assessed scripts of (three) different colleges for detail:

Finally, I have analysed reading & writing skill testing in 34 assessed scripts of three different colleges located in different parts of the country. Among them 13 belong to English First Paper and 21 belong to English Second Paper. In this discussion I will refer to the colleges as college-1, college-2, college-3 and students will be marked as S1, S2, etc., for the sake of the anonymity of the colleges and the markers.

HSC English First Paper:

College-1: Here I am going to examine the reliability of marking mainly, as it is much more problematic. So Q.4, Q.5, Q.7, Q8 & Q13 of English First Paper have been taken into consideration.

It is strange that in spite of writing wrong answers or quoting directly from the passage (by the students), markers are awarding marks, sometimes giving 100% marks, e.g. S3 answers Q8 (making a flow chart). Other examples are S4 – answer of Q8; S5 – answer of Q7 & Q4; though it is a clear violation of marking instruction).

When we studied the answer of Q13 (paragraph writing, we found that almost every answer is memorized, as they do not match with the prompts. 'Obviously a memorized script does not provide an accurate sample of a test-taker's ability…', Weigle (2002:133). The given item is so predictable that students can learn by rote and reproduce it in the exam. It is found that even the content of their writing is alike, which indicates that they learn it by heart from the same source, e.g. S1 answers this question:

And S1 is given 5 out of 14. In total he gets 82 marks for the whole paper. Here we can find *unreliable* marking, as it is clear that a mid-mark is given. This problem also exists in the following reading questions, Q4, Q5, Q7 and Q8 as the students score marks by simply copying from the given passage. For example, the answers of S1.

College-2:

Like the students of College-1, the students of College-2 are also quoting the answers from the given passage when they attempt Q.4, Q.5, Q.7 & Q.8 and most of them obtain around 50% or 50%+ marks. Markers are found to mark whimsically during assessment. For example, S1 writes all answers correctly for Q4 and gets 4 out 5 though according to the marking instruction full credit might be given. S2 & S5 write the same answer with noticeable mistakes and also get 4. On the other hand, S7 attempts the same answer which apparently looks correct but with different formats and he is given 0. We wonder why he gets 0.

We also found that markers put '√' on incorrect answers and putting 'x' on correct answers. Marking is also gravely unreliable in Q13 (paragraph writing). In this answer both S1 & S8 get 5 out of 14. If compared, it will be found that S1 writes a much better answer. S2 & S7 get 7 & 6 respectively. Again a comparative study will find that S2 writes a much better answer. From all these observations the present researcher assumes that a mid mark is given during the assessment of subjective testing. In our observation we found that a marker gives '0' in *rearranging sentences* as a student did not write the sequential number of the sentences. But the same marker gives mark to another student for the same mistake.

College-3 did not provide any scripts.

From the above study it is found that most of the time markers are awarding marks ranging from 80% to 60% while they mark Q.4, Q.5, Q.7 & Q.8. But when they mark Q13, they are unwilling to award more than 45% to 50% marks. It is also established that students get marks as they can easily quote some lines from the reading passage when they attempt Q.4, Q.5, and Q.6 & Q.8. So the *reliability and validity* of testing is in doubt.

HSC English Second Paper:

Like English First Paper I have considered those questions (Q5, Q8, Q11, Q12, Q13) where subjective marking is done. Other questions are objective types (True/False, Right form of verbs, matching column A & B, fill in the gaps, etc.). I have decided to analyse the writing skill testing of English Second Paper as the First Paper has only one writing task (see above).

College-1: Here it is found that students are copying answers from the given passage when they attempt Q5 & Q8. Markers of this college appear to be very strict. It would not have been a problem if these markers would have marked all of the answer scripts of Chittagong Education Board. So inconsistent marking will occur when other markers will mark more liberally than the above markers. Students will be victimised.

It is also found that students produce memorized answers when they attempt Q.11, Q.12, Q. 13 & Q. 14. S1 and S2 produce almost one and the same answers when they attempt Q.13 and Q.14. Moreover, serious low marking is detected almost in every descriptive answer. It is difficult to find reasons behind such low marking. It is also found that markers neither 'x' nor give '0' while they mark. Only an oblique line is drawn which is forbidden in the Marking Instructions. S3:Q13 draws our attention as the student writes an answer which has no relation with the question. We presume that the student memorizes an advertisement which is very common in the HSC test and reproduces it without understanding the question. As a result there is a mismatch.

College-2: Marking is seen seriously inconsistent here as well. When compared the answers of S1 & S2 (Q.12 & Q.13), we find that S1 is given lower marks in spite of writing better answers. S4 also gets better marks though the answers (Q.13 & Q.14) are below standard in comparison with S1. While marking S5 (Q.13) correct writing is unnecessarily underlined and awarded very poor marks. S6 & S8 are awarded poor marks for Q.11Q.12 Q.13 & Q.14.

College-3: In our investigation we find unreliable marking here as well. After comparing the scripts of S1, S2 & S5 (Q5 & Q8) we identified that S2 & S5 are awarded higher marks in spite of producing inferior answers. S1 finally scores 83% marks in the whole paper and so it is expected that the answers will be

good. But we notice that Q13 & Q14 are below standard. The mark given is higher than expected. So the over all reliability of marking is gravely in doubt. That the students are producing memorized answers while attempting Q13 is evident in the scripts of S5, S6 & S7. However, they managed to get 40% to 45% marks.

Findings:

On the basis of the above data analysis, it can now be confirmed that at HSC:

i) Most items do not permit objective scoring. ii) Detail scoring guideline is not provided. iii) Rater's training is not facilitated. iv) Multiple scoring is absent. v) Markers are inconsistent. vi) Examiners are not familiar with the marking system. vii) Candidates are not made familiar with the test format and technique.

So we presume that the ***reliability*** of marking at HSC is problematic.

5. Recommendations:

Here we present the recommendations in the light of the findings of the study.

5.1. Rater training, a burning issue:

White (1984) outlined six practices and procedures which are vitally significant for maintaining high reliability in a large scale assessments. These are as follows:

a) The use of scoring rubrics, b) the use of sample scripts in rater training to exemplify the rating scale, c) double-marking, d) scoring should be in a controlled reading to eliminating unnecessary sources of error variance, e) checks on the reading in progress by reading leaders & f) evaluation and record keeping to differentiate between reliable and unreliable readers.

Considering the constraints mentioned (section-5.2), we can consider at least the first two. We need a formal and regular rater training programme. In training raters, 'it is important to communicate to raters the amount of variability that is acceptable and to let them know that they are not required to be perfectly accurate at all times', Weigle (2002:131). In that training, specific and detailed outline must be formulated regarding the 'off-task scripts', 'memorized scripts' and 'incomplete responses' as the methods of assessing these problematic responses are not mentioned in the rating scale. In addition, 'scoring leaders should be prepared with advice on how to deal with poor hand writing, extremely brief responses, or uncreative or simplistic responses', (ibid:134) .

5.2. Impartial and fair test administration:

The questionnaire data (section-4.1) and reports of national dailies reveal that HSC test administration lacks fair and impartial atmosphere. For the sake of the reliability of the test an absolutely fair and impartial administration must be ensured.

5.3. Rating writing:

During marking raters should give equal importance to the characteristics of the text (content, organization, language use,) and students' abilities (knowledge of grammar/sentence structure). I recommend a marking scale for rating-writing which has been developed by adopting the marking scales of TOEFL (ETS2000) and FCE.

5.4. Test specifications:

Test specifications are the complete guidelines not only for the question setters but also for the test takers and test users as well.

Alderson et al (1995:20-21) justify the necessity of test specification when they write 'the intention of such specifications for candidates should be to ensure that as far as possible, and as far as consistent with test security, candidates are given enough information to enable them to perform to the best of their ability'.

So a complete test specification is essential to make HSC testing more reliable.

5.5. Limitations of the research:

We must take into account some limitations of this research when we are interpreting its results. It is understood from the applied nature of this study that examination hall observations, observation of the markers' orientation programme, interview of examiners, Head Examiners, question setters can be very useful instruments along with questionnaire (cf. McDonough and McDonough: 1997; Nunan: 1992; Hopkins: 2002) in making effective suggestions for increasing *reliability*. As the questionnaire was administered by post, it is not clear how many teacher-examiners had actually received it. However, finally I got 25 out of 59. I also had the difficulty in not having a face to face clarification from the participants, which could have been of greater use.

As HSC is a very large-scale exam, the findings would have been much more realistic and authentic if I could have covered all the Education Boards instead of only one.

5.6. Conclusion:

Despite the limitations, it may well be concluded that the findings of the research provide invaluable information about the *reliability* at HSC Reading and Writing skill tests. Indeed, the relationship between rater training, rating scale and rater reliability is well understood. In addition, fair and impartial test administration is crucially imperative to ascertain the reliability of the test. However, if possible, future study should include more areas of investigation, specially on formulating an exclusive rating scale for assessing reading and writing and effectiveness of rater training in Bangladesh.

Works Cited

Alderson, J.C.; Clapham, C & Wall, D. 1995. *Language Test Construction and Evaluation*, CUP.

Alderson, J.C 2000. *Assessing Reading*, Cambridge: CUP.

Allison, D. 1999. *Language Testing & Evaluation*. Singapore: Singapore University Press.

Anastasi, A. 1988. *Psychological Testing* (6th edition). New York: MacMillan.

Bachman, L. F. & Palmer, A.S. 1996. *Language Testing in Practice*, OUP.

Breen, M., and Candlin. C. N. 1980. *'The essentials of a communicative curriculum in language teaching'*. Applied Linguistics 1(2): 89-112.

Canagarajah, A. S. 2002. "Globalization,Methods,and practice in periphery classroom," *Globalization and Language teaching* (Eds) D.Block and D.Cameron. Routledge.]

Davies, A. 1990. *Principles of Language Testing*, London: Basil Blackwell.

Hoque, S. 2002. English as a second language: A priority programme, *The Daily Star*, January 27, 2002.

Hughes, A. 2003. *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Islam, Z. 2003. *'Bridging the Gap: Curricular Innovation and Teacher Preparation Perspective in Bangladesh.'* Dissertation done in the University of Essex.

Khan, R.A. 2005. Seminar Paper; 3rd International Conference on: "The Effective Teaching of English in Bangladesh: Policy, Pedagogy and Practices". Organized by Department of English. Stamford University, Dhaka, Bangladesh. Fall 2005.

Messick, S. 1989. Validity. In R.Linn (ed.), *Educational Measurement*. New York: Macmillan pp-13-103.

Mathews, J.C. 1985. Examinations: A Commentary. London: George Allen and Unwin.

McDonough, J. & McDonough, S. 1997. *Research Methods for English Language Teachers*, Edward Arnold Ltd., London.

Nunan, D. 1988 a. *Syllabus Design*, Oxford: OUP.

Rea, P. M. 1978: "Assessing Language as communication", *MALS Journal*, Birmingham: University of Birmingham.

Richards, J. C.and Rodgers, T. S 2001. *Approaches and Methods in Language Teaching*. Cambridge: CUP.

Rozario, B J. 2005. "Teaching of English in Bangladesh in Secondary and Higher Secondary Levels," Seminar Paper; 3rd International Conference on: 'The Effective Teaching of English in Bangladesh: Policy, Pedagogy and Practices'. Organized by Department of English, Stamford University, Dhaka, Bangladesh. Fall 2005.

Salim, A & Mahboob T.S. 2001. "ELT and English Language Teachers of Bangladesh: A Profile." *Revisioning English in Bangladesh*. Dhaka: The University Press Ltd.

Shahidullah, M. 2003. Methods of Teaching and Testing English in Bangladesh: the Key Issues, Seminar Paper; Department of English, East West University, Dhaka

organized this seminar on 'English Teaching and Learning in Bangladesh: The Current Issues' (07th August 2003).

Weigle, S. C. 2002. *Assessing Writing*. Cambridge: Cambridge University Press.

Weir, C. J. 1993. *Understanding and Developing Language Tests*. London: Prentice-Hall International (UK).Ltd.

Weir C. J. 2005. *Language Testing and Validation*. London: Palgrave Macmillan.

Newspapers:

The Daily Star <http://www.thedailystar.net/ >

Prothom Alo <http://www.prothom-alo.com/ >

Textbook:

English For Today, For Classes 11-12, NCTB, July2001; Dhaka, Bangladesh.