Reliability of the Currently Administered Language Tests in Bangladesh: A Case Study

Rebeka Sultana
Department of English, University of Information Technology & Sciences (UITS), Bangladesh

Abstract
This study aims at investigating the reliability of the language tests presently in use in Bangladesh, taking the Higher Secondary reading and writing tests as sample. To achieve this purpose, the English question papers of the H.S.C examinations of the last few years have been analyzed with a view to determining the extent to which the test formats used can affect candidates' scores, and involve subjectivity in scoring. Besides, the marking scheme provided to the markers by the education board has been examined in order to find out whether and how far it is able to reduce the element of subjective judgement involved in the assessment of the responses and thus to ensure inter-marker and intra-marker reliability of the tests. Furthermore, an empirical survey has been carried out in the form of questionnaires among the English language teachers of a few colleges in the country to collect quantitative data on various aspects of reliability of the H.S.C language testing. The study reveals that the Higher Secondary reading and writing tests, as a whole, are far below the satisfactory level in terms of reliability as they have deficiency in both test reliability and scorer reliability. Therefore, the scores produced by these tests cannot be considered reliable indicators of the test takers' reading and writing abilities. The shortcomings of these tests, as identified in the investigation, include poor test items, inadequate test instruction, ambiguous marking criteria, insufficient marking guideline, etc. On the basis of the findings, some recommendations have been made for the improvement of the overall reliability of the H.S.C reading and writing tests.

Key Words: Language test, test reliability, scorer reliability, subjectivity, quantitative, marking criteria.

1. Introduction

Language tests play a powerful role in many people's lives, acting as gateways at important transitional moments in education, in employment, and in moving from one country to another (McNamara, 2000). The tests of English at the Higher Secondary level in Bangladesh are one of the most important language tests that play a vital role in the lives of millions of examinees by building up the foundation in English that the students need for higher education in future. The Higher Secondary education comprises only two years, but performance on the tests at this level is the reflection of the candidates' English learning of much longer period, because students in Bangladesh study English language compulsorily for minimum twelve years in their academic life before they sit for the Higher Secondary Certificate (H.S.C) examination. So, it is desirable that those who have passed this examination would be highly proficient in English. Generally, we interpret a score obtained by a test taker as the indicator of the test taker's knowledge or abilities in the particular subject area on which he is tested. But being a teacher in the department of English in a university, I always find that many of my students obtaining a high score in the H.S.C language tests show signs of a very low efficiency in using the language. It seems to be surprising to me that in spite of getting the highest grade (A+) in English in H.S.C, many students at the tertiary level cannot produce even a single sentence in English correctly. This raises some questions in my mind: Why do the test scores obtained by these students fail to reflect the true abilities of the candidates? Is the scoring inaccurate? Or the tests of English themselves are inconsistent in terms of measurement? What steps can be taken to make these tests more reliable so that they provide consistent, replicable information about the testees' language performance, and make meaningful distinctions among students in terms of the abilities being measured? We all know that the importance of English language competency can in no way be denied in today's competitive world. It is also clear to us that a reliable language test at the H.S.C level can go a long way towards building up a bright academic career of a large number of students by enabling them to cut a good figure in the subsequent admission tests, and to successfully pursue higher studies in which most of the authentic books are written in English. So, it is obvious that finding out the answers to these questions is highly significant in the present context of our country. In search of the answers, therefore, I have chosen the H.S.C English tests as the object of my research focusing specifically on the tests of reading and writing skill.

2. Theoretical Background

2.1 Reliability

Reliability is an absolutely essential quality of any good test, which refers to the consistency of measurement, and the stability of test scores (Harris, 1969). If a test is administered to the same candidates on different occasions, then, to the extent that it produces different results, it is not reliable (Heaton, 1976). Similarly, if the same test papers are marked by two or more different examiners, or the same examiner on different occasions, then, to the extent that different marks or grades are awarded, it is not reliable (Heaton, ibid). It is clear from the foregoing that two different types of consistency or reliability are involved here: reliability of the test itself, and reliability of the scoring of the test. Reliability is thus divided into two types: test reliability, and scorer or rater reliability (Harris, ibid). It is important, as Harrison (1983) points out, that the student's score should be the same or as nearly as possible the same whether he takes one version of a test or another, and whether one person marks the test or another. The more similar the scores would have been, the more reliable the test is said to be. Scorer reliability is further subdivided by Hughes (2003) into two types: intra-scorer reliability and inter-scorer reliability. Hughes (ibid) defines 'intra-scorer reliability' as the reliability of one person scoring the same test responses on different occasions, and 'inter-scorer reliability' as the reliability of different people scoring the same test responses.

Heaton (ibid) discusses five factors that can affect the reliability of a test:

i) **The extent of the sample of material selected for testing:** The larger the sample, i.e. the more tasks the testee has to perform, the greater the probability that the test as a whole is reliable.

ii) **Fluctuations in test administration:** Test reliability is adversely affected if the conditions under which the test is administered lend to fluctuate from one administration to another.

iii) **Personal factors:** Poor health, fatigue, lack of interest or motivation, and emotional disturbance of the examinee can lower the reliability of a test.

iv) **Test instructions:** Test instructions play a crucial role in test takers' performance, since their performance depends, to a large extent, on how well they understand the procedures to be followed and the nature of the tasks they are to complete.

v) **Fluctuations in scoring:** Subjectivity in scoring may introduce inconsistencies in scores and produce unreliable measurement.

The amount of time allocated for the test or its parts is also a factor concerning the reliability of a test. Madsen (1982, cited in Bachman, 1990) mentions inadequate time allocation as a source of test anxiety, and hence, influence on test performance. There is evidence in literature that test performance is also influenced by the particular test format used. McNamara (2000) identifies two types of test format: i) fixed response format in which the candidate's task is to choose the appropriate response from those offered, ii) constructed response format in which the response consists of a production of a language sample in response to the input material. Numerous research studies (for example, Clifford, 1981; Bachman and Palmer, 1981; Shohamy, 1984) have demonstrated that test format has a sizable influence on performance on language tests. Fixed response formats, for example, are found to result in guessing which has a considerable but unknowable effect on test scores. Thus, the scores gained in multiple-choice and true/false tests may be suspect because the candidates have guessed all or some of the answers (Weir, 1998). Bachman (1990) emphasizes that if we want the test scores to be an accurate estimate of the candidates' language abilities, we clearly need to undermine the detrimental effects of the techniques of measurement on candidates' performance.

Hughes (2003) makes a number of suggestions on how to make tests reliable. He says that reliability can be achieved by: i) taking enough samples of behaviour, ii) excluding items which are unable to sort out better candidates from the poorer ones, iii) imposing restriction on the candidates' choice of tasks, iv) providing brief, simple and unambiguous test directions, v) making sure that tests are well laid out and perfectly legible, vi) making candidates familiar with the testing techniques, vii) providing a supportive test-taking environment to *all* the candidates, viii) using items that permit scoring as objective as possible, ix) providing a detailed scoring key to the scorers, x) employing multiple independent scoring where testing is subjective.

2.2 Objective and Subjective Testing

'Subjective' and 'objective' are terms used to refer to the scoring of tests. In an objective test, the correctness of the test taker's response is determined entirely by predetermined criteria so that no particular knowledge or training in the examined content area is required on the part of the scorer (Henning, 1987). A common example would be a multiple- choice recognition test. Conversely, a subjective test is said to be the one in which the scorer must make an opinionated judgment about the correctness of the response based on his subjective interpretation of the scoring criteria (Henning, ibid). An example might be the scoring of free, written compositions for the presence of *creativity* in a situation where no operational definitions of *creativity* are provided and where there is only one rater. In general, the less subjective the scoring, the greater agreement there will be between two different scorers, and between the scores of one person scoring the same test paper on different occasions. Therefore, the term 'subjective' is often used to denote *unreliable* or *undependable*. However, subjective tests can be objectified in scoring through the use of precise rating schedules clearly specifying the kinds of errors to be quantified, or through the use of multiple independent raters (Henning, ibid).

2.3 Approaches to Scoring

Two basic approaches to scoring students' writing samples are identified in literature (Hughes, 2003): i) Holistic and ii) Analytic. Holistic scoring (often referred to as 'impressionistic' scoring) involves the assignment of a single score to a piece of writing on the basis of a total impression of the writing as a whole. This kind of scoring has the advantage of being rapid. Experienced scorers can, as Hughes (ibid) finds, judge a one-page piece of writing in just a couple of minutes or even less using a holistic scale. But holistic evaluation would appear to be more ubjective as it depends on the impressions formed by the markers (Weir, 1998). According to Hughes (2003: 95), however, "if well conceived and well organized, holistic scoring in which each student's work is scored by four different trained scorers result in high scorer reliability".

Analytic scoring, on the other hand, requires separate score for each of a number of components of a performance (e.g., relevance and adequacy of content, compositional organization, cohesion, paragraph, grammar, choice of vocabulary, mechanical accuracy, etc.). Even where analytic rating is carried out, it is usual to combine the scores for the separate aspects into a single overall score for reporting purposes (McNamara, 2000). This method of scoring has the advantage that it can provide a vivid profile of students' strengths and weaknesses in different aspects of writing (Weir, 1998). On behalf of analytic scoring, Hughes (ibid) claims: i) it disposes of the problem of uneven development of sub-skills in individuals, ii) it compels the scorers to consider aspects they would otherwise ignore, iii) The very fact that the scorer has to give a number of scores will tend to make the scoring more reliable. Moreover, an analytic mark scheme is seen as a far more useful tool for the training and standardization of new examiners. Francis (1977, cited in Weir, 1998) points out that, by employing an analytic scheme, examining bodies can better train and standardize new markers to the criteria of assessment. There is, however, a disadvantage associated with the analytic method of scoring. As teachers look at specific areas, marks are often lower than that may be achieved by using holistic scale (Jacobs et al.,1981).

3. Research Methodology

This study has used three different sources for collecting data, and involved three steps: a) Analysing the question papers (English First Paper & Second Paper) of the H.S.C. examination, prepared by eight general education boards (Dhaka, Rajshahi, Jessore, Comilla, Chittagong, Sylhet, Barisal and Dinajpur) from 2011-2014, b) Analysing the scoring criteria given in the instructions to the examiners, c) Collecting the views of 20 EFL teachers from six colleges (Govt. Hazi Mohd. Mohsin College, Ctg; Chittagong University College, Chittagong Govt. Women College, Chittagong Islamia University College, Chittagong Laboratory College, and Women College, Enayetbazar, Chittagong) in Chittagong city about different aspects of the H.S.C language tests by means of questionnaire.

Murphy (1979, cited in Weir, 1998) helps us in drawing up a list of questions on the basis of which the appropriateness and effectiveness of the marking scheme and the tasks set for the candidates can be evaluated. Using Murphy as the informing source in this regard, a checklist has been prepared for the analysis of the question papers and marking instructions (Appendix-A).

This study is basically a descriptive study. Random sampling method has been used to select respondents for the questionnaire survey, and quantitative method is used to collect data for this study. As the teachers selected for the survey have a long-term experience of working as both invigilators and examiners, they have been able to provide rich data on test reliability and marker reliability of the present H.S.C testing. The questionnaires distributed among them consist of six closed questions related to the aspects of reliability.

4. Question Format

The H.S.C language tests consist of two compulsory papers. Two hundred marks are allotted for the two papers, one hundred marks to each. Paper I is divided into three parts: 'Seen Comprehension', 'Vocabulary' and 'Guided Writing'. Marks allocated to these three parts are 40, 20 and 40 respectively. Paper II is divided into two parts: 'Grammar' and 'Composition' carrying 40 and 60 marks respectively. For each paper, students are allowed three hours for completing the tasks.

5. Analysis of the Question Papers

In the existing test format at the Higher Secondary level, 40 marks have been allocated for the reading test in which candidates are asked to read short extracts taken from a set textbook and then to perform various types of tasks based on the given extracts. The testing techniques include multiple-choice questions, true/false questions, gap filling, short-answer questions, information transfer, and summary writing. Five multiple-choice items are given each having three options. The advantage of using this technique is that in this type of tests there is almost complete marker reliability, since candidates' marks, unlike those in subjective formats, cannot be affected by the personal judgement or idiosyncrasies of the marker (Weir, 1998). But the fact that the responses on a multiple-choice test (a, b, c, d) are so simple makes them easy to communicate to other candidates non-verbally and thus cheating may be facilitated (Hughes, 2003). Another objection to the use of multiple-choice tests is that candidates can learn strategies for taking such tests that 'artificially' inflate their scores: techniques for guessing the correct answer, for eliminating implausible distractors, for avoiding two options that are similar in meaning, and so on (Alderson et al.,1995). According to Heaton (1976), guessing can be a serious factor in influencing scores in multiple-choice questions containing only three options. Heaton, therefore, suggests setting at least four or five alternatives for each item in order to reduce the possibility of guessing. An analysis of the multiple-choice items found in the H.S.C question papers further reveals that some of these items can easily be answered without reference to the texts they are set on as candidates are likely to be able to answer them from general knowledge without even reading the text, and if this is so, whatever it is that is being tested, it cannot be comprehension of the texts. Thus, the MCQ format is unable to make any distinction between the better students and the poorer ones, and therefore lacks test reliability. Items in which the test taker has merely to choose between TRUE and FALSE are nothing but a variety of multiple choice with only two options, and a 50% chance of choosing the correct response by chance alone (Hughes, 2003). Five questions of this type are given carrying 5 marks. Many educators feel that true/false items serve little or no measurement purposes. Heaton (ibid), however, assures that the scores obtained by the testees on a true/false test can be reliable indices of reading comprehension if there are a lot of items. Here the true/false questions set for the H.S.C candidates have been found to be so simple and easy that they do not at all suit the stage and the standard of the learners. The next task is to fill in five gaps with clues, and five without clues. A possible weakness of gap-filling items, as identified by Hughes (2003), is that it is difficult to make sure that there is only one correct answer for each gap. The scoring of these items, however, can be highly reliable if it is carried out with a carefully constructed key in which there are more than one answer for some spaces so that the scorers can rely on it completely and do not have to use their individual judgement (Hughes, ibid). In contrast to the three items discussed here, the short-answer questions provide more reliable data about the test takers' reading ability (Hughes, ibid). The H.S.C candidates are asked to answer five short-answer questions which require them to supply, as opposed to select, a response. The purpose of this kind of item is to elicit a relatively unconstrained response, which may vary in length from a few words to an extended paragraph. The longer the required response, the greater the difficulties of scoring of this kind (Hughes, ibid). In those cases where there is debate over the acceptability of an answer, e.g., in questions requiring inferencing skills, there is a possibility that the variability of answers might lead to marker unreliability (Weir, 1998). The next item is information transfer which requires the candidates to show successful completion of the reading task by presenting information given in the text in a different format. The H.S.C question paper sets two types of information transfer tasks for the examinees. One is to make a list from the ideas contained in the passage, and the other to transfer materials from the given text on to a flow-chart. Information transfer minimizes demands on candidates' writing ability, and avoids possible contamination from students having to write answers out in full (Weir, ibid). The last task is to summarize the main points of the given extract in five sentences. The inclusion of an integrated writing component of this type as part of any test is again problematic from the marking point of view. To assess the responses reliably one needs to formulate the main points contained in the extracts, construct an adequate mark scheme and standardize markers (Weir, ibid). Identifying the main points in a text is itself so subjective that the examiners may not agree as to what the main points are (Alderson et al.,1995).

The writing test is divided into two parts. Part I is 'Guided Writing' carrying 40 marks, which consists of three types of tasks. Firstly, candidates are asked to match some phrases in a substitution table to make sensible sentences. They have to write the sentences out in full. This item tests students' understanding of sentence construction; it does not require them to produce any written language. Secondly, candidates are presented with a series of mixed-up sentences and are asked to write them in the correct order to make a coherent paragraph. This task also does not involve writing; it only tests students' awareness of 'internal reference', i.e., the way in which words and phrases, particularly pronouns, relates to each other in a paragraph or text. So, the scoring of these two items is quite straightforward and, therefore, reliable. The third task is to write a paragraph answering some specific questions, which demands subjective decisions from the scorer.

Part II is a composition test carrying 60 marks and consisting of five different tasks. In this test, the test takers are given topics for writing a report, a short composition, an application and a dialogue, and are given some clues (two or three lines) to complete a story. To score a composition test reliably and consistently is highly difficult, but the great advantage of using this type of tests is that they can measure certain writing abilities more effectively than do objective tests (Harris, 1969). With regard to the importance of the subjective tests, Heaton (1976) says, it is impossible to obtain any high degree of reliability by dispensing with the subjective element and attempting to score solely on an "objective" basis. Some frequently used topics for the H.S.C candidates for writing short compositions are 'Students and Social Service', 'Wonders of Modern Science', 'Benefits of Reading Newspaper', etc. Some common issues of writing reports include 'The rising of prices of essential commodities', 'A village fair you have visited', 'A massive fire on a garment factory you have experienced'. According to Heaton (ibid), in a composition test, the candidates should be presented with a clearly defined problem which motivates them to write. The tasks should be such that they ensure the candidates have something to say and a purpose for saying it, and also an audience in mind when they write. This is supported by Weir (1998) who opines that when the tasks are determined precisely by specifying for each of the tasks: the media, the audience, the purpose and the situation in line with target level performance activities, it becomes easier to compare performances of different candidates and to obtain a greater degree of reliability in scoring. But here we find that the writing tasks that are given neither clarify the purpose of the writing nor do they provide any meaningful context for writing. This type of free, uncontrolled writings give the examinees a lot of scope to cover up weaknesses by avoiding problems they find difficult (Harris, 1969). Here it is also noticed that in some of the question papers, candidates are offered a choice of writing either a dialogue, or a summary of a given passage, and are given three topics from which to choose one for writing short composition. If candidates are given the freedom to choose items from a number of alternatives, it becomes difficult to ensure that all candidates have undertaken equivalent tasks (Harris, 1969). Moreover, no instruction is given (in English Second paper) on the expected length of answers of these writing tasks. It seems that the test writers have decided to rely on the students' powers of telepathy to elicit the desired behaviour. Another problem with these tasks is that they do not explain the assessment criteria; candidates are only asked to write but they are not given any idea about how their writing is going to be marked. Are the markers looking for fluency or accuracy? Are the marks awarded for the structure of the composition, and the ability to present a good argument, or solely for the use of grammar and vocabulary? Candidates need to know all these things in order to decide whether to use easy, well-known structures so as not to be penalized for errors, or whether to take risks because extra marks are awarded for the use of complex and creative language (Alderson et al.,1995). Since the necessary information are all lacking in the instructions in the question paper, it is very likely that some candidates would put emphasis on the length of writing, whereas others on appropriacy of style; to yet another, clarity of argument would become more important than all other criteria. Reliability of scoring of the H.S.C writing test has been considerably affected by these problems.

6. Analysis of the Marking Scheme

Murphy (1979: 19, cited in Weir, 1998) outlined the nature of the marking scheme demanded by the Associated Examining Board: 'A marking scheme is a comprehensive document indicating the explicit criteria against which candidates' answers will be judged: it enables the examiner to relate particular marks to answers of specified quality'. Every year 'The Board of Intermediate & Secondary Education' in Bangladesh provides the examiners with a marking scheme entitled 'Instructions to Examiners' for evaluating the answer scripts of the board examinations. For the purpose of the current research, one of such documents provided by the Chittagong board in 2012 has been collected. This includes instructions for marking both Paper I and II, but our discussion here will be limited to the marking of only the reading and writing components. Some problems have been found in these instructions which are discussed below:

I. English First Paper: Instruction-f says, 'All mistakes- ***spelling, grammatical or otherwise*** should be underlined with red ink/ coloured pencil' (Appendix B-1). This instruction makes the measurement of the abilities in question less accurate. In this regard, Hughes (2003) argues that in a reading test, errors of grammar or spelling should not be penalised, provided that it is clear that the candidate has successfully performed the reading task which the item set. He goes on to say that the function of a reading test is to test reading ability; to test productive skills at the same time which is what happens when grammar, etc. are taken into account simply makes the measurement of reading ability less valid. Similarly, overemphasis, according to him, on such mechanical features as spelling and punctuation also invalidates the scoring of written work.

II. **a. English First Paper:** Instruction-i says, 'answers having ***originality, creativity and grammatical accuracy*** should be given maximum credit for question no. 7 and 13 (Appendix B-1). Question no. 7 refers to the summarizing of the reading passage, and question no. 13 to the writing of paragraph based on the given questions.

b. English Second Paper: Instruction no. 11 states, **'*originality, creativity and grammatical accuracy*** should be specially weighed' (Appendix B-2). This instruction is to be used for marking the short composition of the writing test.

c. English Second Paper: Instruction no. 14 says, 'title is a must in developing the story. ***Originality and creativity*** should be given special credit (Appendix B-2).

The three instructions above seem to be somewhat puzzling. In these instructions candidates' creativity and originality have been considered to be the most important rating criteria. Here it seems that the problem in instruction arises due to problem in the setting of questions. Because, as Hughes (ibid) says, in language testing the testers should not be interested in knowing whether students are creative, imaginative, or even intelligent.

Harris (1969) gives the same advice: the testers must avoid setting tasks that require a high degree of ingenuity and creativity. For Harris (ibid) the purpose of general writing-ability testing is to elicit characteristic samples of every student's writing and from these to determine his proficiency at expressing himself in clear, effective and grammatical prose - not to measure his "creative powers". Heaton (1976) distinguishes between the terms 'composition' and 'essay' by making it clear that while writing an essay involves far more than the production of grammatically correct sentences and demands creativity and originality, the writing of a composition should involve the students only in manipulating words in grammatically correct sentences and in linking those sentences to form a piece of continuous writing which successfully communicates the writers thoughts and ideas on a certain topic. Good essayists, he argues, are as rare as good poets since essays are intended not only to inform but also to entertain and impress. So, in a composition test, while it is reasonable, according to Heaton (ibid), to expect the learner to write accurate English for a meaningful purpose, it is neither reasonable nor realistic to demand originality and creativity in the form of an essay as it would be like requiring a poet to sit down and compose an original poem in half an hour or so under examination conditions. So, it is obvious that the marks allocated to the tasks in this composition test are not at all commensurate with the demands the tasks make on the candidates.

There is a high degree of subjectivity, as we have seen, in the scoring of items of the composition test. Henning (1987) finds that any rater called upon to make subjective estimates of composition quality in a language is liable to be inconsistent in judgement. He also observes that raters sometimes differ widely among themselves in estimates of appropriate marks for the same sample of language performance. In both cases, inconsistency occurs in scoring, and this particularly happens, as Henning (ibid) points out, in situations where the raters are not provided with detailed rating schedules. Here we find that no rating scale is prescribed for the raters, and no explicit criteria are there which the raters adhere to in marking the subjective items. Even no guideline is found in the 'Guided Writing' section (Appendix B-1) for marking the paragraph. Candidates' marks in the writing test is thus compelled to be largely affected by the particular examiner who assesses them. A scoring key is provided for the reading test, which allows for a few possible alternative answers for the gap-filling items. But we have seen that there are also some items in the reading test, which call for short written responses, and can confront the examiners with a variety of possible acceptable answers. Since the marking scheme does not perfectly specify performance criteria, the element of subjective judgement that the examiner has to exercise in evaluating candidates' answers on these items is not reduced at all. Moreover, as the mark scheme does not indicate clearly the marks to be awarded for the relative weighting of criteria that might be applicable, it can be easily interpreted by a number of different examiners in a different way. Thus it fails to ensure that all the examiners mark to the same standard.

From the above discussion, it is clear that the instructions are not self-explanatory, rather some of them are ambiguous which make them totally unreliable, especially for the marking of the extended writing tasks, and since the scoring of the tests is not reliable, we can say without any doubt that the test results are not reliable either.

7. Results of the Questionnaire Survey

The findings of the empirical survey are presented below:

The Teachers' Responses to the Closed Questions

Sl. No.	Questions	Responses			
		Yes	%	No	%
1.	Are the candidates able to perform the tasks satisfactorily in the time allowed?	19	95%	1	5%
2.	Do you follow any marking guideline while marking answer scripts?	15	75%	5	25%
3.	Do you think marking varies significantly from examiner to examiner?	17	85%	3	15%
4.	Do you find it necessary to employ more than one scorer for reliable scoring?	14	70%	6	30%
5.	Do you feel the necessity of a rating scale for the assessment of the written production?	17	85%	3	15%
6.	Is uniformity maintained in all administrations of the tests?	2	10%	18	90%

In the above table, question no 1 and 6 are intended to elicit data on test reliability whereas the rest are concerned with the issues of marker reliability. The responses to the question asking about the allotment of time indicate that the tests do not put any undue pressure on the examinees with regard to the time allowed. But the response to the question on the administration of the tests is indicative of an inadequate test reliability of the current H.S.C tests as 90% of the respondents have said that all examinees are not allowed to perform under identical testing conditions. The responses to question no 3 and 5 clearly indicate that one of the major sources of inaccuracy of the tests is the variation in scores caused by the application of different standards by different markers. From the responses to question no 5, it is clear that there is also inconsistency on the part of the individual markers as 85% teachers are found to express the need for an analytic scheme which can facilitate agreement amongst examiners as to the precise range of qualities that are to be evaluated in a composition. Finally, it can be said that the findings in this survey make the rater reliability of the testing at H.S.C questionable.

8. Overall Findings

The findings of the study can be summarized as follows:

Both fixed response format and structured response format have been used in the H.S.C English testing. Therefore, the assessment of some of the responses is 'objective' while the assessment of some others is 'subjective'. The fixed response items used permit completely objective scoring, and possess a high degree of scorer reliability, but these items are unable to discriminate widely enough among the testees. They keep a lot of scope for the weaker students to copy in the examination hall and easily obtain marks which they do not deserve. Strong students and weak students can perform with similar degrees of success on these items. Moreover, the tasks in the reading test are not set at an appropriate level of difficulty. Some of them are too easy compared to the level of Higher Secondary students. These easy, non-discriminating items have been contributing very little to the reliability of these tests.

The 'constructed response' items used are found to require the production of a number of writing samples varying in length from a single sentence to an extended piece of discourse. Thus, some test items lie somewhere between the extremes of objectivity and subjectivity whereas some others are fully dependent on the scorers' subjective judgement. The examiners score the papers of both internal and board examination by adopting an impressionistic method of marking, and a single marker is employed for each script. So, it is clear that scoring in the prevailing testing system at H.S.C is liable to make the markers extremely unreliable both in their own inconsistency and in their failure to agree with the other examiners on the relative merits of a student's composition. It is found that most of the markers follow some guidelines for marking, but the problem is that the instructions for marking themselves are so defective, confusing and insufficient that it is useless whether the teachers follow them or not; these instructions can never ensure equal marks for equal levels of performance.

The examinees are allowed a reasonable amount of time to complete the tests, but are not provided with the necessary stimulus and information required for writing. Some carelessly worded instructions are found to be used to elicit samples of writing, which can be interpreted in various ways by the test takers, leading to non-equivalent performances. Thus the vague writing assignments make it easier for the test takers to conceal their inadequacy in language ability, and make it difficult for the testers to compare performances of different candidates. The data provided by the teachers reveal that different testing situations in different examination centres are also responsible for variation in performances.

9. Recommendations

The following recommendations emerge from the systematic investigation to help the H.S.C language tests give us a fair measurement of the examinees' language performance:

1. Items like multiple choice, true/false, etc. which may involve guessing on the part of the test takers and encourage cheating in the examination hall and which do not discriminate well between weaker and stronger students should either be excluded from the test format, or a large number of items of similar kind should be added to the existing items so that the test method effect is eliminated, or at least minimized to some extent, and test reliability is increased. If multiple-choice items are to be used there should be at least five options for each item. The distractors, or incorrect options should be reasonably attractive and plausible in each item and should appear right to the testees who are unsure of the correct option. Care should be taken to ensure that it is not possible to answer correctly purely on the basis of outside knowledge or to eliminate some of the choices because they are clearly illogical or because they conflict with one another. The true/false items can be modified by instructing the test takers to give a reason for their choice.
2. Sufficient information should be conveyed by the rubrics in a simple language written in a concise and lucid manner in order to provide a realistic, helpful basis for writing. The free writing items should be replaced with controlled writing tasks with efficient, effective instructions in which a well-designed *prompt* outlining a purpose and context for the tasks is given. Thus the tasks should be specified so that the test takers have less freedom in the way that they respond. In order for the test takers to have the opportunity to perform at their best, they should also be provided with information about the exact nature of the testing procedure and the test tasks. Specially, it is crucial for the instructions to mention the desired length of the responses, and state explicitly the criteria that will be used for evaluation.

3. The range over which the possible answers might vary should be restricted. So, candidates should not be given a choice of items. In order to facilitate marking, comparison across the candidates should be made as direct as possible by making *all* the candidates perform the same tasks and write on the same topics.
4. For assessing the reading tasks, a scoring key should be provided to the scores which specifies acceptable answers and assign points for acceptable partially correct responses. For high scorer reliability the key should be as detailed as possible in its assignment of points. It should be the outcome of efforts to anticipate all possible responses and have been subjected to group criticism.
5. Question setters should be given proper training on language testing issues so as to give them better exposure to the techniques of constructing more reliable test items.
6. Introducing multiple marking or even double marking of scripts of the board examination may not be feasible since large numbers of scripts are involved and the results are to be published within a short period of time. In that case, an analytic rating scale should be developed for the raters so that both intra-rater and inter-rater error variance in rating can be reduced.
7. Since scoring of the tests necessarily involves subjectivity on the part of the scorers, every scorer should be trained in the application of the scoring criteria through rigorous standardization procedures.
8. A uniform and non-distracting condition of administration should be ensured in all teaching institutions of the country.

10. Conclusion

Due to some limitations, the procedures followed by the test administrators in administering language tests in Bangladesh could not have been extensively surveyed in this study. The study would have been more holistic if the administrative processes involved in testing could also be included in the area of investigation. Here it is felt that a further study should be conducted in future, if possible, investigating the issues of test administration, with a detailed survey on the extent of the administrators' interaction with the examinees, the prevention of cheating in the examination centre, the adequacy of the personnel involved in the tests, the physical characteristics of the test environment, such as–temperature, humidity, seating arrangement, lighting, etc. Another limitation of this research project is that it has not been possible in this study to cover a large area of the country for empirical survey as it would have involved a lot of time, money and manpower. Therefore, the sample size has been kept rather small and confined to only six colleges. Despite these limitations, the present study has successfully revealed many of the drawbacks of the H.S.C language tests, which are responsible for the failure of the tests to accurately measure the abilities that they are designed to measure. Implementation of the recommendations is expected to make the tests maximally useful in providing information about the testees' abilities in reading and writing by removing the drawbacks.

Appendix A

Checklist for the Analysis of Question Papers and Marking Instructions

1. Are the tasks unambiguous, giving a clear indication of what the examiner is asking, so that no candidate may take the task to mean something different?
2. Are the tasks too easy or too difficult for the candidates taking the tests?
3. Can the tests discriminate between students of varying ability?
4. Do the complexity and length expected of the candidates adhere to the amount of marks the questions carry?
5. Does the mark scheme anticipate responses of a kind that candidates are likely to make?
6. Does the mark scheme indicate clearly the marks to be awarded for different parts of a question or the relative weighting of criteria that might be applicable?
7. Does the mark scheme allow for possible alternative answers?
8. Does the marking scheme, by specifying performance criteria, reduce as far as possible the element of subjective judgement required from the examiner?
9. Can the marking scheme be easily interpreted by different examiners in the same way?

Appendix B-1

BOARD OF INTERMEDIATE & SECONDARY EDUCATION, CHITTAGONG

Email: info@bise-ctg.gov.bd, Website: www.bise-ctg.gov.bd

Higher Secondary Certificate Examination, 2012

Subject: English (Compulsory) First Paper

INSTRUCTION TO EXAMINERS

1. Examiners are requested to adhere to the Board's rules for guidance as enunciated in their appointment letters.
2. Examiners are requested to be aware of putting their signatures and Code No. neatly on every script, mark

sheet and in the space provided for the purpose.

3. Examiners must award marks in English figures both inside the script and on the OMR. They are requested to prepare mark sheets in English as well. The prescribed Proforma supplied by the Board must be duly filled in and sent accordingly with each installment of scripts.

Marking of Scripts:

a. Marks of different answers should be written on the cover page against specific question numbers showing the division of marks, if any.

b. Awarding of marks in fraction should be avoided as far as possible. If the total number on the cover page is a fraction, it must be turned into the next whole number.

c. If an examinee writes more than the required number of answers to a question, the one considered better should be marked and the rest should be cancelled by making comment 'Excess' on the margin.

d. Marginal marks from 28 to 32 in any script should be re-examined.

e. Answers deserving no credit should be given zero (0) and not simply be (X) crossed.

f. All mistakes - spelling, grammatical or otherwise should be underlined with red ink / coloured pencil.

g. No credit should be given to irrelevant answers and if it happens so, the word 'irrelevant' should be written on the margin. Not more than pass marks should be given to partially relevant answers of question no: 5 and 7.

h. Not more than 50% marks should be awarded for memorized answers. Unintelligent, crammed and bad reproduction should be given less credit.

i. Answers having originality, creativity and grammatical accuracy should be given maximum credit for question no: 7 and 13.

j. Any marginal marks that may deprive an examinee of his getting grade point should be avoided.
38, 39→40; 48,49, →50; 58,59, →60; 68,69, →70; 78,79, →80.

k. No extension of time beyond the last date fixed by the Board will be allowed. If an examiner fails to finish his / her work within the scheduled time, he must return the unexamined scripts with relevant papers to the Head Examiner without any delay.

Probable Answers for Some of the Questions of H.S.C. English Exam. 2012, Paper I

Reading Comprehension

Q. No.1. Examinees are supposed to write only the right words. a) impression b) persuaded c) debates d) father e) stay

Q. No.2. If any examinee writes only 'False', without giving correct information, he/she should be awarded not more than 50% marks.

a) True
b) False. Correct Information: Rabindranath liked to listen to the debates of the parliament.
c) False. Correct information: London created a poor/ bad impression on young Rabindranath.
d) False. Correct information : Tagore's bother was in Brighton.
e) False. Correct information : Mr. Scott's daughters didn't like him from the moment they saw him.

Q.No.3. If any examinee writes the correct form of words without maintaining the right preposition, he/ she should be awarded 50% marks. No spelling mistakes can be considered in any case.

a) wrote, b) arrived in, c) listened to, d) paying, e) foggy.

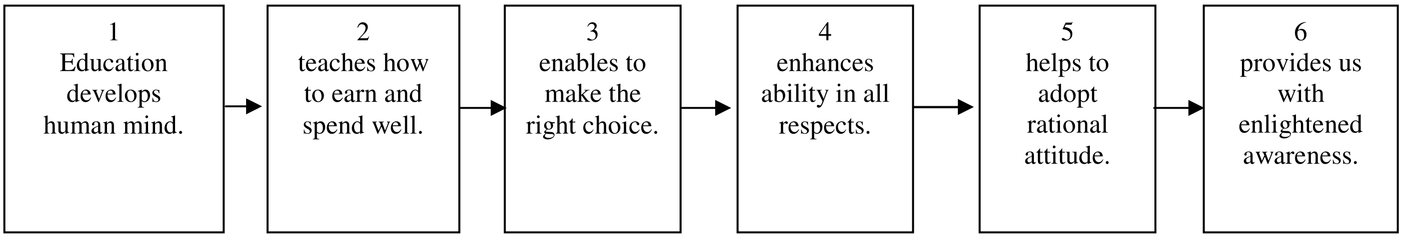
Q. No.4. Answers should be written in words/ phrases/ sentences.

Q. No.5. Each question should be answered in brief, to the point and in a single sentence.

Q. No.6. a) regarded / thought / deemed / taken
b) having / receiving / getting
c) free
d) removes / eradicates
e) awareness / consciousness / enlightenment.

Q.No.7. Summary has to be treated as a holistic idea. Answers produced verbatim from the original passage should not be awarded more than 40% marks.

Q.No.8. The necessary information has to be presented in the form of words/phrases/sentences in a diagram horizontally or vertically or in any form maintaining the sequence.



1 Education develops human mind.

2 teaches how to earn and spend well.

3 enables to make the right choice.

4 enhances ability in all respects.

5 helps to adopt rational attitude.

6 provides us with enlightened awareness.

Guided Writing

Q.No.-11. I) Khan Jahan Ali found Bagerhat beset with various problems.
II) He built numerous mosques in Bagerhat.
III) The Shat Gambuj Mosque is the most magnificent of them.
IV) It was used both as a prayer hall and a court of Khan Jahan Ali.
V) The mosque is regarded as one of the architectural beauties of the country.
VI) The UNESCO had declared the mosque as a World Heritage Site.

Q.No.-12. II - V- I- VI- XI- X- XIII- VIII- IX- III- XII- IV- VII- XIV
or
II–V–VIII– X–XIII– IX– I –VI – XI –III –XII– IV–VII–XIV.

Any logical and suitable answer should be given credit.

Appendix B-2

BOARD OF INTERMEDIATE & SECONDARY EDUCATION, CHITTAGONG

Email: info@bise-ctg.gov.bd, Website: www.bise-ctg.gov.bd

Higher Secondary Certificate Examination, 2012

Subject: English Second Paper (Compulsory). Code No. 108

Specific Instructions

Composition

Q.No.10.Answer to the given question should be written under the required title in the form of reports.

Q.No.11.Originality, creativity and grammatical accuracy should be specially weighed. Memorized answer should not be given more than 60 % marks.

Q.No.12. Due credit should be given to the answer having form and contents. Mere form without content should not be awarded any credit.

Q.No.13. Dialogue developed in communicative English should be given special credit.

Q.No.14. Title is a must in developing the story. Originality and creativity should be given special credit.

References

Alderson, J.C., Clapham, C. and Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L. F. and Palmer, A. S. (1981). 'The construct validation of the FSI oral interview.' *Language Learning* 31,1: 67-86.

Clifford, R. T. (1981). 'Convergent and discriminant validation of integrated and unitary language skills: the need for a research model' in Palmer *et al.:62-70.*

Francis, J.C. (1977). 'Impression and analytical marking methods', memeo, MS Aldershot: Associated Examining Board.

Harris, D. P. (1969). *Testing English as a Second Language.* NY: McGraw-Hill.

Heaton, G.B. (1976). *Writing English Language Tests.* Second edition. London: Longman.

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation, Research.* Cambridge, MA: Newberry House.

Hughes, A. (2003). *Testing for Language Teachers.* Second edition. Cambridge: Cambridge University Press.

Jacobs, H.L., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F. and Hughey, J.B. (1981). *Testing ESL composition: a practical approach.* Rowley, Mass: Newbury House.

Madsen, H. S. (1982). 'Determining the debilitative impact of test anxiety.' *Language Learning* 32: 133-43.

McNamara, T. F. (2000). *Language Testing*. Oxford: Oxford University Press.

Murphy, R. J. L. (1979). *Mode 1 examining for the General Certificate of Education. A general guide to some principles and practices,* mimeo, Guilford: AEB.

Shohamy, E. (1984). 'Does the testing method make a difference? The case of reading comprehension.' *Language Testing* 1,2: 147-70.

Weir, C. J. (1990). *Communicative Language Testing.* New York: Prentice Hall.