IIUC STUDIES
ISSN 1813-7733
Vol.- 6, June 2010 (p 35-52)

Necessity of initiating Rating Scale for more reliable assessment of *writing skill at HSC level: a case study*

Mohammed Humayun Kabir*

Abstract: *In this essay, firstly, I will discuss importance of* ***rating scales****, relationship between* ***assessment criteria*** *and* ***operations and conditions*** *and effectiveness of rating scales while assessing writing. Secondly, I will examine a selection of scales and subsequently there will be an estimation of those in meeting objectives of the tests and of the course. Finally, I will recommend a suitable rating scale to test English First paper and English Second Paper writing skill at HSC level in Bangladesh evaluating the existing one where I will mention steps to be taken to reduce inter and intra- rating fluctuation in scoring.*

Rating scale:

A rating scale is *'an ordered set of descriptions of typical performances in terms of their quality, used by raters in rating procedures'* McNamara (2000:136). Hudson at el (1995) utilizes rating scales *'to assess test takers' pragmatic performance by external raters and for test takers' self-assessment of their own pragmatic ability*' , cited in Hinkel (2005:328). Its aim is to make distinction sufficiently fine *'to capture progress being made by students'*. It is a realistic way of assessing the level of a particular communicative performance *'by using a number of descriptive bands for a particular skill, on a scale of competence ranging from excellence to failure'* Nunn (2000:171). For the sake of a reliable assessment we need a reliable rating scale. As a rating scale decides a range of things including scoring validity, students' competence or proficiency, raters should be trained up to standardise marking.

* Assistant Professor, Department of English Language & Literature, IIUC

Rating scales and their effectiveness in meeting purpose:

There are mainly two types of rating scales- a) Holistic / Global scale b) Analytic scale.

Holistic / Global scale:

'*Holistic scoring (sometimes referred to as 'impressionistic scoring') involves the assignment of a single score to a piece of writing on the basis of an overall impression of it'* Hughes (2003:94) e.g. TOEFL test. A 'band scale' is used in these sorts of marking. It has both advantages and disadvantages.

Advantages:

It is a response to writing as a whole that emphasises on *'what is done well, not on problems'*. This sort of marking is not time consuming. In public examination where there are millions or thousands of examinees the concerned authority has to choose this type of scale.

Disadvantages:

It lacks diagnostic information. 'Scores generated holistically cannot be explained to the other readers in the same assessment community; diagnostic feedback is out of the question' Hamp-Lyons (1995:759). One possible disadvantage of holistic judgment is that *'different raters may choose to focus on different aspects of the written product'* Nakamura (2002). As it awards only one score, it reduces reliability.

Analytic scales:

'Methods of scoring which require a separate score for each of a number of aspects of a task are said to be analytic' Hughes,(2003:100) e.g. TEEP. Here marks awarded for these aspects 'are based on some form of text analysis rather than general impression'.

Advantage:

These scales *'guard against ignoring aspects thought to be important'* Lilley (2007), as they provide diagnostic information. A notable advantage of analytic scoring '*is that raters are required to focus on each of various assigned aspects of a writing sample , so that they all evaluate the same features of a student's performance'* Nakamura

(2002). These scales are easier for training purposes and useful for wash back.

Disadvantages:

It is a lengthy process of marking and expensive as well. It might have *'halo effect'*. In its application there might arise some issues regarding judgements which may be difficult to make. Here scores may not be informative.

Therefore, we can conceive that no scale is flawless. None of the above mentioned scales can guarantee an absolutely reliable judgement. They are not cent percent effective in meeting the purpose. Appropriate criteria need to be drawn up to improve the reliability of assessing writing. These criteria should be relevant to the operations and conditions specified (see estimation of different rating scales in the latter part of the discussion). Training in their interpretations and use is critically important; otherwise reliability will be seriously compromised. So a suitable marking scale, criteria and level of marking, descriptors, raters ' training experience of markers, etc. are vitally important in order to make a rating scale effective in meeting the purpose.

Importance of Rating scales:

The importance of rating scale is felt greatly specially where there is subjective assessment. Not only the assessors but also the test takers need rating scales. It provides the students with *'a realistic goal by describing the performance just above her or his present level'* Nunn (2000:171). The main general advantage of designing rating scales for any course is the harmony that can be achieved between the potentially discordant and conflicting perspectives of teachers, learners, and assessors. Alderson (1991:71-85) discusses the reasons for using rating scales in some detail, but only a short outline will be presented here. Firstly, rating scales provide an easily understandable report (op.cit:72) for candidates, administrators, course designers and teachers on the level of performance of individuals or groups, at the same time by providing descriptions of what candidates can do .They can report on 'typical or likely behaviours of candidates at any given level or on the proportions of candidates at each level. Secondly, rating scales can guide the rating process (op.cit.73) standardizing the criteria for an individual rater or act as 'a common standard for different raters'. Finally, they also help to guide the construction of task (op. cit: 74) which allow students to display the described behaviours at their own levels [cited in Nunn (2000)].

In short, we presume that rating scales help assessors to decide what level or score is to be awarded to each test taker in a test. Underhill (1987:98) states rating a scale *'offers the assessors a series of prepared descriptions, and she then picks the one which best fits each learner'*. Of course, a well developed rating scale can contribute significantly to ensuring a more reliable measurement.

When assessors assess a writing task and follow a particular rating scale they should have a thorough idea about the operations ,conditions , and assessment criteria for the test because these fundamental considerations are directly linked with the assessment of a particular task. A comprehensive idea about these elements will facilitate a rater to follow a rating scale more accurately.

RELATIONSHIP BETWEEN ASSESSMENT CRITERIA AND OPERATIONS AND CONDITIONS:

Before establishing the relationship between assessment criteria and operations and conditions for the testing of writing I include a short discussion on operations and conditions and assessment criteria.

Operations:

These refer to *'the tasks that candidates have to be able to carry out'* Hughes (2003:60). Hence it is these are the ways in which '*test information helps us to make decisions about the test participants'* Carroll and Hall (1985:101). Operations are divided into two levels;- i) macro level and ii) micro level. In macro levels there can be found two categories: a) Interactional and b) Informational.

Interactional activities are personal letter writing, creative writing, expressing thanks, opinions, apology, justification, complaints, etc.

Informational activities are particularly applicable to academic environment. These include describing and defining phenomena, describing process, and giving instructions, argumentation and critical evaluation.

At HSC level test takers have to carry out some of the above mentioned interactional informational activities in their English examinations **(see below).**

On the other hand micro linguistic level of operations can also be specified. These include grammar, vocabulary, punctuation, handwriting, etc.

Conditions:

These contain a number of psycholinguistic considerations i.e. text type (genre) required, topic, audience, time available, amount of support given, familiarity with task type, stated or un-stated criteria of assessment, etc.

Text type:

Validity and reliability of writing test have been found to be increased by sampling more than one task/text type appropriate to the students writing needs. *'The more samples of a student's writing in a test, the more reliable the assessment is likely to be'* Weir (1993:134). *'In general it is felt advisable to take at least two samples'* Jacobs et al. (1981:15).

So More than one samples are appreciated for testing. When we examine the tasks of both English 1st and 2nd papers, we will find at least four types of text.

Topic:

Raimes (1983:266) has strongly recommended that choosing topic be the teachers' most responsible activity. *'It is necessary to ensure that students are able write something on the topic(s) they are presented with'* Weir (1993:134). In the task '*common background information'* needs to be provided. It should be appropriate and realistic to the students' needs. Students should write on the same topic and preferably more than one sample of their ability should be measured. According to Jacobs et al (1981:1), '*it is generally advisable for all students to write on the same topics because allowing a choice of topics introduced too much uncontrolled variance into the test'*.

In writing test HSC test takers do not have any option when they choose a topic .All students have to endeavour same task. However, there are lots of problem with the ***common back ground information***.

Amount of time allowed for each task/size of output:

These are very practical and realistic issues for testing as well as real-world. Sufficient time should be allowed so that students can produce a coherent text. It should be **long enough** to be marked reliably. *'If we want to establish whether a student can organise a written product into a coherent whole, length is obviously a coherent factor,*' Weir (1993:135). '*Both time and length need to be stated'*. However, Kroll (1990) reports that there is little significant difference in quality of writing done either under time pressure or over a longer period (in terms of language and organizational skills).

It should be mentioned that '*care needs to be taken to provide clear instructions and an idea of assessment criteria'* Lilley (2007).

Assessment Criteria:

Raters are to assess the *'learner output in terms of overall levels of performance, as demonstrated in or inferred from the task products*' Allison (1999:175). While assessing a writing task a marker should look into relevance and adequacy of the content, organization, cohesion, vocabulary, grammar, punctuation, spelling, etc. as assessment criteria. Raters can get the idea about the assessment criteria from the prescribed rating scale.

Relationship:

From the above discussion it is evident that there is a deep and inseparable relationship between operations, condition and assessment criteria. What functions are going to be expressed through writing tasks (operations) is directly related with text type/genre, time available, length (conditions).And assessment criteria decide the level of performance of a performer considering the quality of operations and conditions .So all three are interactive. *'The absence of one reflects badly on the others'* Weir (1993:136).

Estimation of some famous rating scales:

In this stage I am going to estimate three rating (two global and one analytic) scales for marking writing papers. There I will discuss how far these are effective in meeting the objectives of the test/ course.

Firstly, I am going to examine TOEFL writing scale as one of the most recognized holistic making scales.

TOEFL score is used to evaluate English proficiency of people by the college or university administration in making decision about admission purpose .This score is also used by Govt. agencies, scholarship programmes, etc. as well to choose potential candidates. In this testing 30 minutes time is allocated for writing section which consists of a single essay. No choice is given and the scale that is used for scoring task is highly structured. TOEFL bulletin (ETS 2000) states that the purpose of writing test is *'to demonstrate [test takers'] ability to write in English. This includes the ability to generate and organise ideas, to support those ideas with examples or evidence, and to compose in standard written English in response to an assigned topic'* (p-41). All TOEFL writing prompts are disclosed in the most current TOEFL bulletin .One of them is selected randomly by computer. These prompts are of two types .In one type students are *'to express and support an opinion*' and in the other type they are to *'choose and defend a position on an issue'*.

Scoring:

A six- point holistic rating scale is used to score TOEFL writing test (See Appendix-i). It addresses the following aspects of writing : *'overall effectiveness of the response to the writing tasks'* , how far it is organised and developed, use of details , *' facility with the use of language'* and syntactic variety and appropriate word choice. A smaller number of qualified raters mark the scripts working from an ETS (Educational Testing Service) established scoring centre or from their homes using a web interface. *'Raters have access to sample essays, both keyed and handwritten, on all topics at all score points and work under the supervision of Scoring Leaders who monitor the performance of these raters in real time and can contact them and be contacted while scoring is taking place*' Weigle (2002:145). It should also be mentioned that all raters must pass a calibration test at the beginning of each scoring day before they start scoring. A *'reported* *score is the average of the two raters' scores'* and if there is *'a discrepancy of two points or more an experienced rater marks it'* and *'the final score is the average of two closest scores'*(ibid). In this test reliability of scoring is gained *'through careful pre-testing of prompts and rigorous training and monitoring of raters* (ibid)*'*.

Limitations:

It does not have different '*audiences*' or '*purposes*'. So, there the construct being measured is limited to a narrow focus. Furthermore there *'seems to be an implicit bias towards privileging linguistic accuracy over other aspects of writing such as task fulfilment and development in TOEFL writing test'* Weigle (2002:146). The authenticity of this testing is limited as there no scope '*to read about or discuss the assigned topic before writing about it'* and test takers are not allowed to choose prompts. In order to increase the interactiveness a choice of prompts may be offered to examinees.

Cambridge First Certificate in English (FCE):

FCE marking scale is used to assess English language ability for office work or to pursue a training course in English. It consists of five dissimilar papers and the second paper is the writing paper. Two writing tasks (one is obligatory and the other one is elective task) are included in writing paper. The compulsory task is a 'transactional letter' and the optional task is to be done by choosing one out of four (a task on the reading of one of the five books specified in advance, and other tasks cover a variety of genres like non-transactional letters, discursive compositions, narratives and descriptions) .Examiners are to produce each of the tasks between 120 and 180 words within 1hour 30 minutes.

Scoring:

These writing tasks are scored on a six –band scale-*'a general impression scale*' (See Appendix-ii). A *'task specific mark scheme'* is drafted in advance of each test administration and *'is finalized after consideration of actual written samples'*. Part-1 insists on organizing of the coverage of content points, while the range of structures and vocabulary used is indicative of performance in part –2. Scores 'are are converted to provide a score out of 20 for each piece of writing'. These pieces of writing are assessed by a panel of trained examiners. The examiners are divided into small teams headed by a Team Leader.

The Principal Examiner guides and watches the marking process. A common standard of assessment is maintained by the selection of each sample scripts for all the tasks on the writing paper. There is no double –rating system but *'a rigorous process of co-ordination and checking by Team Leaders is carried out before and during the marking process, and procedures for examiner-scaling are in place in order to minimize subjectivity'* Weigle (2002:151). 'Pass' or 'Fail' is not decided in a particular paper, 'but rather in the examination as a whole'. The three passing grades are –A,B,C, and two failing grades are D and E. Statement of results as well as graphical display of performance in each of the five test papers is given to the examinees.

The use of different rating scales and the different tasks may give *'a true picture of the test takers' range of abilities'* (ibid). It has also increased the authenticity of the test .As the examinees can respond to the task they feel best equipped to handle, it ensures much interactivity.

Limitations:

Since two test takers may score similar grade through very different means, it is difficult to say accurately what it is that FCE is testing. Furthermore, the wide range of tasks *'may detract from reliability'* of the test.

Test in English for Educational Purpose (TEEP):

The criteria in marking scheme 2 in TEEP *'resulted from a survey of a large number of academic staff at tertiary- level institutions in the United Kingdom'*. The academic staffs expressed their opinion in favour of the procedures that *'would assess students, particularly in relation of their communicative effectiveness'*. They also sought a profile *'containing details of candidates' strength and weakness'*. In this test *'the candidates have to extract specified information from an article provided'* Weir (1993:162). Here all the '*lexis is provided for the candidates, either through labelled diagrams or in available text*' (ibid).

In this scale there are provisions for separate scores for each of the several criteria. The TEEP scales presented by Weir (1993:156) cover seven criteria: relevance and adequacy of content, compositional organization, cohesion, adequacy of vocabulary for purpose, grammar, mechanical accuracy- i (punctuation), and mechanical accuracy-ii (spelling).

Scoring:

Here each of the criteria is sub-divided into four behavioural levels on scale of 0-3(See Appendix-iii). A level 3 a base line of minimal competence and at this level it was felt that a student was likely to have very few problems. At level 2 a limited number of problems are found and at level-1 a test taker needs a lot of help, while level-0 indicates almost total incompetence in respect of the criteria in question. I can see that the two mechanical accuracy criteria were considered to be of very little importance and I think serious thought might be given to omitting them in future.

Limitations:

It is evident that first four levels are related to communicative effectiveness but the latter three emphases accuracy. *'It may well be that the latter three criteria contribute to communicative effectiveness or lack of it, but attempts to incorporate some indication of this proved unworkable*' (ibid: 162). Another problem is noticed as all the lexis is provided for the candidates 'the likelihood is that most candidates will score reasonably well on the adequacy of vocabulary criterion'.

In the above discussion we have evaluated 3 world class popular rating scales comprising both analytic and holistic scales .Considering the practical issues we feel that we need a holistic rating scale for assessing HSC writing skill .The discussion made below will focus on the types of tasks at HSC and will be followed by a proposed rating scale.

Writing skill testing at HSC:

Here I am going to describe rating procedure used at HSC level in Bangladesh. The HSC English First Paper contains the following types of writing tasks under the heading ***'Guided Writing'*** : i) producing sentences from substitution tables. ii) re-ordering sentences and
iii) Answering questions in a paragraph.

40 marks have been allocated for these three (task-i-12 marks, task-ii-14 marks, task-iii-14 marks) types of writing out of 100 marks for this paper (See Appendix-iv).

In task-i students are given a table containing clauses and phrases in different columns in an unorganised way. Students have to make six meaningful sentences from this given table. If the sentences are correctly formed but sequence is not maintained full marks might be awarded. This task may be different in nature, e.g. students might be asked to form questions from answer supplied or they might be asked to complete a dialogue where some questions and some answers are supplied and some are missing. Students provide the missing bits. As this is a writing test spelling or punctuation errors are penalised. Though there is a provision for different patterns of tasks for testing writing skills, I have never seen that those are set up in the previous test papers. That is why students generally prepare themselves for the former type of task.

In task-ii students are to re-order or re- arrange 14 sentences where they have to maintain sequence. If the sequence is broken with any sentence or sentences marks are deducted for that sentence or sentences. However, marks are given for all other sentences arranged sequentially .The third task is answering questions in a paragraph. Here a topic related to a student's own experience is given which contains four to six questions. This task is designed in such a way as it has a reflection of real-life, communicative contexts.

In English Second Paper 40 marks allocated for Reading skill(form unseen reading comprehension passage) ,20 marks for Grammar & 40 marks for Writing skill testing. (**Appendix**-v**).**

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]**

Scoring:

Here subjective marking is done. Raters are advised in the Teachers' Guide to take into account the following criteria during assessment:

'markers therefore need to look at two distinct criteria :a) task fulfilment i.e. has the student satisfactorily answered what he has been asked to do? b) The mechanics of good writing i.e. are the ideas well organized? do the paragraph develop logically? are sentences within a paragraph and across paragraphs linked in a cohesive manner? etc'. (see appendix-vi).

In reality we find that markers hardly follow Teachers' Guide during marking. Khan (2005), an ELTIP (English Language Teaching Improvement Project) teacher trainer and researcher points out that , '*teachers are over burdened with heavy work loads with little time to spare for lesson planning class preparation or correctiomn of written work. Access to* ***teachers' guide is nil****…'*

It is surprising that the marking instructions for assessing English 1st and 2nd paper writing parts (Appendix-vii) do not possess any specific assessment criteria. **In the instructions for English 1st paper** it is directed that examiners should not give more than 75% marks(for **Q11,Q12)** if any student write (paragraph/composition) without a title. Neither the T.G. nor the syllabus for HSC textbook has mentioned it. This is a very sensitive and crucial issue directly related to the interest of the test takers and test – users. Why is it not well circulated? Why are only markers informed about this? It is as if the education board authorities 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, -*'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'***. Moreover, there is not any instruction about marking the Question No.13 which is only writing item in English 1st paper.

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. If the instruction is obeyed there is every possibility that students will be tempted to memorize answers of the probable questions. These answers will vary from students to students, 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.

It should be noted that there is no separate marking for writing skill in the final marks transcript of the student(See Appendix-vi) .Students are awarded a Grade (A+,A,A-,B,C,D,F) making a total of that s/he scores in reading test, Vocabulary test and Writing test. So we see that global rating scale is followed here.

A suitable marking scale, criteria and level of marking, descriptors, raters' training, experience of markers, etc. are vitally important in order to make a rating scale effective in meeting the purpose.

Considering time constraints, resource crisis, financial factors above all huge number of examinees (more than half a million) our Education Ministry so far has not initiated any analytic scale. Hence, so far at HSC a holistic scale is followed.

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).This will also have the role in establishing validation of the test.

It is true that training only is not enough and it cannot guarantee a valid and reliable score. To make sure fair, consistent and reliable marking we need to monitor the marking done by the markers.

Limitations:

Here it is noticed that the operations of the test are inappropriate. If we carefully observe the tasks, we will find that other than the paragraph (Q.No.13) writing students need not to produce any creative task. Students are just tailoring the bits or re-arranging them in the 1st and 2nd task. It is very surprising that the only real writing task has strictly word limitation. Students are to produce the answer within 100words. In fact, we cannot deny that the length of task is an important criterion to judge the student's writing ability.

In the 2nd paper students are to attempt four types writing tasks. As the scoring process is purely subjective and there exists no clear marking instruction we assume serious unreliable marking as Kabir (2000) finds in his research.

Recommendations:

The operations should be much more interactional (letter writing, creative writing, etc.) and informational (describing process, argumentation, critical evaluation, etc.) in line with CLT. If it is so, students will be compelled to practice more types of tasks and it will help them to develop their writing skill to a great extent. Furthermore, students will not be interested to memorize answers of the probable questions. As in Bangladesh students get their final results in a single transcript which contains marks of 12 papers (subjects), as is an exam where millions of test takers take part, as there is an extreme crisis of raters and no provision for **inter- rater rating**, as the results are to be published within three months, we have to rely on a global scale. We agree '*in the assessment of writing , a major advantage of holistic over analytic scoring is that each writing sample can be evaluated quickly by more than one rater for the same cost that would be required for just one rater to do the scoring using several analytic criteria(cf.* Davies et al, 1999)' cited in Nakamura(2002). Students may be asked to write a paragraph which will be lengthy enough to judge their ability of communicative competence as well as language skill. It is also an important issue of **assessment criteria** and **reliability of scoring**. While marking raters can give equal importance to the characteristics of the text (content, organization, language use,) and students abilities (knowledge of grammar/sentence structure), adopted from Upshur and Turner (2002). If special attention is given to these criteria we can significantly ensure **intra- rater reliability.**

Rating scale for assessing writing:

I recommend a marking scale that I proposed in my research done in the University of Essex, UK, (Kabir2000). It has been developed by **adopting the marking scales of TOEFL (ETS2000) and FCE**, as assessing the writing parts at HSC is found very problematic in the data analysis. It addresses the following aspects of a piece of writing :overall effectiveness of the response to the writing task, organization and development of idea, facility with use of language, syntactic variety, diction and task fulfilment, etc. It is presented below:

Recommended marking scale for HSC writing test:

Marks to be awarded out of 10	**Marks to be awarded out of 14**	**Criteria /descriptors**
8 – 9	10 –12	*efficiently addresses the writing task, *well organized and well developed *displays consistent facility in use of language *demonstrates syntactic variety and appropriate word choice though it has occasional errors *no sign of producing memorized

		answers / copying from question paper. Overall: a very positive effect on the target reader.
6 –7	8 –9	*addresses the topic adequately but may omit slight parts of the task *adequately organized and developed * demonstrates adequate but possibly inconsistent facility with syntax and usage*may contain some errors that occasionally obscure meaning *some evidences of producing memorized answers / copying from question paper. Overall: a positive effect on the target reader.
4 –5	6 –7	* Inadequate organization or development * a noticeable inappropriate choice of words * errors in sentence structure and/or usage * little or no detail, or some irrelevant specifics * evidences of producing memorized answers / copying from question paper. Overall: message not clearly communicated to target readers.
0 –3	0 –5	*serious disorganization or underdevelopment * serious and frequent errors in sentence structure or usage *serious problems with focus, or no attempt at answer * evidences of producing crammed answers / copying from question paper. Overall: a negative effect on the target reader.

(Adopted from the marking scales of TOEFL (ETS2000) and FCE marking scales.)

Conclusion:

In this essay I have described the importance of rating scales and its types including operations ,conditions and assessment criteria with special reference to assessing writing .I have also estimated TOEFL and FCE as holistic and TEEP as analytic scales. Last of all I have discussed the HSC English 1st and 2nd Paper rating procedure and proposed accordingly a feasible rating scale so that reliable assessment might be ensured.

References:

ALDERSON (1991), '*Bands for Scores'* in Alderson and North (1991)

ALDERSON, J.C.; CLAPHAM, C. AND WALL, D. (1995), *Language Test Construction and Evaluation*. Cambridge University Press.

ALLISON, D. (1999), *Language Testing & Evaluation*. Singapore University Press.

CARROLL, B. J. AND HALL, P. J. (1985), *Make Your Own Language Tests.* Pergamon Institute of English Pergamon Press, Ltd.

ETS (2000): Test of English as a Foreign Language (TOEFL). Princeton, NJ: Educational Testing Service.

HAMP-LYONS, L. (1995), *'Rating Non-native Writing: The Trouble with Holistic Scoring'* TESOL Quarterly 29/1995.

HINKEL, E.(EDS) (2005), *Handbook of Research in Second Language Teaching and Learning*. Lawrence Erlbaum Associates' publishers. London.

HUGHES, A.(2003), *Testing for Language Teachers*. Cambridge University Press.

JACOBS, H.L.; S.A. ZINKGRAF; D.R. WORMUTH; V. FAYE HARTFIEL AND J. HUGHEY (1981), English Composition Program. *Testing ESL Composition: a Practical Approach*. Rowley, Mass.: Newbury House.

KABIR, M. H (2007) : '*An Investigation into the Validity and Reliability in Testing Reading* and Writing skills at HSC level in Banglades*h'.*

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.

KROLL, B.(ED) (1990), *Second Language Writing*. Cambridge University Press.

LILLEY, T. (2007), LG-666,WK-21, Hand out

MCNAMARA, T. (2000), *Language Testing*. Oxford University Press.

NAKAMURA, Y. (2002), *'A comparison of holistic and analytic scoring methods in the assessment of writing'.* The Interface Between Inter language., Pragmatics and Assessment: Proceedings of the 3rd Annual JALT Pan-SIG Conference.

NUNN, R. (2000), '*Designing rating scales for small group interaction'*. ELT Journal 54/2.

RAIMES, A. (1983), *'Anguish as a Second Language?* Remedies for Composition teachers'.

In A. Freedman et al (eds.), *Learning to Write: first language/second language*. London. Longman.

UPSHUR, J. AND TURNER, C (2002): *'Rating Scales Derived From Student Samples:Effects of the Scale Marker and the Student Sample on Scale Content and Student Scores.'* TESOL QUARTERLY, VOL-36. NO-1.

UNDERHILL, N.(1987), *Testing Spoken Language*. Cambridge University Press.

WEIGLE, S. C. (2002), *Assessing Writing*. Cambridge University Press.

WEIR, C. J. (1993), *Understanding and Developing Language Tests*. Prentice Hall International (UK). Ltd.