International Journal of Applied Linguistics & English Literature
ISSN 2200-3592 (Print), ISSN 2200-3452 (Online)
Vol. 5 No. 6; November 2016
Australian International Academic Centre, Australia
Flourishing Creativity & Literacy

A Study of the Validity of English Language Testing at the Higher Secondary Level in Bangladesh

Chowdhury Mohammad Ali
Department of English, University of Chittagong, Bangladesh
E-mail: cmali1958@yahoo.com

Rebeka Sultana (Corresponding author)
Department of English, University of Information Technology & Sciences (UITS), Bangladesh
E-mail: rebeka_uits@yahoo.com

Received: 16-03-2016 Accepted: 29-07-2016 Advance Access Published: September 2016
Published: 01-11-2016 doi:10.7575/aiac.ijalel.v.5n.6p.64 URL: http://dx.doi.org/10.7575/aiac.ijalel.v.5n.6p.64

Abstract

Validity is considered to be of paramount importance in language testing, and therefore, remains the central concept to all designs and research activities in the field of testing and assessment. Arguably, all researches in language testing are in some senses about validity and the process of validation. In this regard, it is the intent of the present research to investigate the validity of the English language tests employed at the Higher Secondary level in Bangladesh. The research questions addressed concern finding out whether the tests are valid in terms of content and construct. The tests administered at this level are 'achievement tests', designed to measure the extent of learning in a prescribed content domain in accordance with explicitly stated objectives of a learning program. The first objective of the study is, therefore, to examine how far the course objectives are reflected in the contents of the existing tests. Secondly, the study makes an assessment of how well these tests measure the abilities they are intended to measure. The findings reveal a great mismatch between what the tests aim at testing and what they actually test. A wide gap is found between the curriculum goals and the existing test format. The study also finds that the Higher Secondary language tests are largely unable to measure the constructs they are based on. The key recommendations to increase the content and construct validity of these tests include developing test specifications and designing syllabus in accordance with course objectives, using direct tests and authentic tasks, sampling widely and unpredictably, arranging training programs for the language teachers, etc.

Keywords: Validity, achievement test, test specifications, syllabus, direct test, authentic task.

1. Introduction

Of the many issues involved in testing and assessment, validity in particular has always been of major concern to all testers. It has been identified as 'the most important quality to consider in the development, interpretation, and use of language tests' (Bachman, 1990:289). The present study aims to evaluate the prevailing English language testing system at the Higher Secondary level in Bangladesh in terms of validity. Language tests are taken mandatorily by the students studying at this level, which are one of the major public tests used with a large test population throughout the country. Tests can be classified into various types according to the purposes they serve. The Higher Secondary language tests fall into the category of 'achievement tests'. This type of tests are 'directly related to language courses, their purpose being to establish how successful individual students, groups of students, or the courses themselves have been in achieving objectives' (Hughes, 2003:13). So, the first consideration in evaluating these tests is the adequacy with which the test contents can fulfill the test objectives. At the same time, the appropriateness of the syllabus in terms of the stated purpose of the course has also been considered, since these tests are deliberately constructed as a sample of the syllabus and materials. Next, the study examines how successful the tests are in measuring the skills they claim to measure.

2. Research Problem

A large number of studies pertaining to test validity can be found in literature, but no significant research has been conducted so far on this issue in the context of Bangladesh. Although Kabir (2009) has attempted to carry out a research within the area selected for the present investigation, many important aspects of validity have been overlooked in his study, which demand consideration. The conceptions of construct validation, for example, have not been discussed, and no idea is given about the constructs underlying the acquisition of the macro skills. Moreover, there is no discussion on the characteristics of communicative testing. Most important of all, the recommendations made do not seem to be sufficient to overcome the existing problems. The present study is designed with a view to filling up the gap, and to provide a deeper insight into the current language assessment situation at the Higher Secondary level in Bangladesh.

3. Theoretical Framework

3.1 Validity

'Validity' in language testing has traditionally been understood to mean discovering whether a test 'measures accurately what it is intended to measure' (Hughes, 1989:22), or uncovering the 'appropriateness of a given test or any of its component parts as a measure of what it is purposed to measure' (Henning, 1987:170). Heaton (1975:153) defines the validity of a test as "the extent to which it measures what it is supposed to measure *and nothing else*". Harris (1969) defines validity with reference to two questions: "(1) What precisely does the test measure? And (2) How well does the test measure?"

To use test wisely we need information about what types of inferences can reasonably be made from test scores. This is a matter of validity, which "refers to the **appropriateness, meaningfulness** and **usefulness** of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences" (APA 1985:9). In order to support or justify the inferences we make about the quality or qualities of the test takers, we must first clearly define the construct, and then we need to develop an argument that the test, the test tasks, and the test scores are relevant not only to the construct but also to the test purpose (Douglas, 2010). Thus, the notion of validity raises the question of the extent to which the score is relevant and useful to any decisions that might be made on the basis of scores, and whether the use of the test to make those decisions has positive consequences for test takers (Fulcher, 2010). The question of relevance and usefulness relates to whether it can be shown that the inferences we draw from a test score about the knowledge, skills and abilities of a test taker are justified (Fulcher, ibid.) If a test is not valid for the purpose for which it was designed, then the scores do not mean what they are believed to mean (Alderson et al., 1995). So, if we claim that a test provides information on a number of different skills or abilities, it should be structured and scored according to the skills and abilities of interest (Fulcher, ibid).

According to Henning (1987), the term 'valid', when used to describe a test, should usually be accompanied by the preposition *for.* Any given test then may be valid for some purposes, but not for others. The matter of concern in testing is to ensure that any test employed is valid for the purpose for which it is administered. Whether the test is for use in the classroom, or for large-scale administration, we need a convincing argument that it is useful for its purposes (Kane, 2006).

3.1.1 Content Validity

Content validity is concerned with whether or not the content of a test is sufficiently *representative* and *comprehensive* for the test to be a valid measure of what it is supposed to measure (Henning, 1987). A test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned (Hughes, 2003). 'Content validity is the *representativeness* or *sampling adequacy* of the content - the substance, the matter, the topics - of a measuring instrument '(Kerlinger, 1973: 458). Anastasi (1982:131) defines the term as: 'essentially the systematic examination of the test content to determine whether it covers a representative sample of the behaviour domain to be measured'. If the test construction process involved appropriate methods for defining the original domain, and writing items that sample that domain in an appropriate fashion, then the test should have adequate content validity (Gifford, 1989). From the standpoint of content validity, an ideal test would be one which required candidates to perform all the relevant potential tasks (Hughes, 2003).

Anastasi (1982:132) provides a set of useful guidelines for establishing content validity:

i. 'the behaviour domain to be tested must be systematically analyzed to make certain that all major aspects are covered by the test items, and in the correct proportions';
ii. 'the domain under consideration should be fully described in advance, rather than being defined after the test has been prepared';
iii. 'content validity depends on the relevance of the individual's test responses to the behaviour area under consideration, rather than on the apparent relevance of item content.'

This kind of validity depends on a careful analysis of the language being tested and of the particular course objectives. Tests should be so constructed as to contain a representative sample of the course, the relationship between the test items and the course objectives always being apparent (Heaton, 1975). Content validation also involves analyzing the content of a test and comparing it with a statement of what the content ought to be. Such a content statement may be the test's specifications, it may be a formal teaching syllabus or curriculum, or it may be a domain specification (Alderson et al., 1995). A test's specification is the blueprint to be followed by test and item writers, which provides the official statement about what the test tests and how it tests it (Alderson et al., ibid). It is a detailed document, and is sometimes confidential to the examining body. The content validity of a test is assured by the accuracy of the test's specification (Harrison, 1983). Deriving from a test's specifications is the test syllabus.

3.1.2 Construct Validity

"Construct validity concerns the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of abilities, or constructs" (Bachman, 1990:255). "If a test has construct validity, it is capable of measuring certain specific characteristics in accordance with a theory of language behaviour and learning" (Heaton, 1975:154). For example, if the assumption is held that systematic language habits are best acquired by means of the structural grammar approach, then a test which emphasizes the lexical or situational meaning of language rather than the structural meaning will have low construct validity (Heaton, ibid). To measure the construct validity of a test, a

tester must articulate the theory underlying his or her test, and then compare the results with that theory (Alderson et al., 1995).

Ebel and Frisbie (1991:108) give the following explanation of construct validity:

> *"The term* construct *refers to a psychological construct, a theoretical conceptualization about an aspect of human behaviour that can not be measured or observed directly. Examples of constructs are intelligence, motivation, anxiety, attitude, dominance, and reading comprehension. Construct validation is the process of gathering evidence to support the contention that a given test indeed measures the psychological construct the makers intend it to measure. The goal is to determine the meaning of scores from the test, to assure that the scores mean what we expect them to mean".*

The term 'construct validity' is therefore used to refer to the extent to which we can interpret a given test score as an indicator of the abilities, or constructs, we want to measure (Bachman and Palmer, 1981).Thus, construct validity pertains to the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores (Bachman and Palmer, ibid). The greater a test's content validity, the more likely it is to be an accurate measure of what it is supposed to measure, i. e. to have construct validity (Hughes, 2003). The extent to which a test is successful in measuring what it sets out to measure also depends largely on the effectiveness of each of the items used (Heaton, 1975).

Two major sources of threats to test validity are worth noting: 'construct irrelevant variance' and 'construct under-representation' (McNamara, 2000). In the threat to validity known as 'construct irrelevant variance', the assessment is too broad containing many variables which are irrelevant to the interpreted construct. McNamara (ibid) gives an example- the knowledge or skill being tested may be embedded in a context which is neither within the candidate's experience nor relevant to the thing being assessed. In an oral test, candidates may be asked to speak on an abstract topic; if the topic does not match their interests or is one about which they may have little knowledge, the performance is likely to appear less impressive than when the candidates are speaking about a more familiar topic at an equal level of abstraction. In this case, then, a potential problem is that the trait being assessed, i. e. ability to discuss an abstract topic in the foreign language, is confounded with the irrelevant requirement of having knowledge of a particular topic. By contrast, in other cases, the assessment may be deficient; the test may be too narrow and may fail to include important dimensions or facets of focal construct. The extent to which a test does not measure the relevant constructs is the degree to which it under-represents the constructs that are supposed to be assessed (Fulcher, 2010).

3.2 Communicative Language Test

Communicative language tests are used with the goal of assessing language learners' ability to use language for communication in specific contexts, involving productive language either through meaningful input for the test taker to comprehend or interpret, or as meaningful output generated by the test taker (Douglas, 2010).

The communicative paradigm, as it is sometimes called, was developed in part in response to an earlier approach to language teaching and testing, the *structuralist* approach, which involved analysing the language into its component parts– phonemes, morphemes, syntactic forms, etc. and assessing them separately, often without reference to context of use or communicative purpose. The testing of separate individual points of knowledge in this way was known as discreet point testing (McNamara, 2000). The philosophy behind the communicative approach was that even if a learner knows all the bits and pieces of a language – the sound system, the vocabulary, the grammar – he would still be incapable of communicating effectively. What the learner needs in addition to language knowledge is *communicative competence*, or the ability for language use (Hymes, 1972), which involves judgements about what the grammar will allow one to say and about what is socially appropriate to say in a given situation. Communicative competence, as Canale and Swain (1980) specified, consist of three components:

i) **grammatical competence** (knowledge of systematic features of grammar, lexis and phonology);

ii) **sociolinguistic competence** (knowledge of rules of language use in terms of what is appropriate to different types of interlocutors, in different settings, and on different topics); and

iii) **strategic competence** (the ability to compensate in performance for incomplete or imperfect linguistic resources in a second language)

In 1983 Canale updated this model by subdividing socio-linguistic competence, which still relates to socio-cultural rules, but he introduced a further competence, that of discourse. **Discourse competence** concerns mastery of cohesion and coherence in different genres.

Morrow (1979) felt that a distinction needed to be made between communicative competence and communicative performance, the distinguishing feature of the latter being the fact that performance is the realization of Canale and Swain's (1980) three competences and their interaction: 'in the actual production and comprehension of utterances (under general psychological constraints that are unique to performance)'. Morrow (1979) and Canale and Swain (1980) argued that communicative testing, as well as being concerned with what the learner knows about the form of the language and about how to use it appropriately in contexts of use **(competence),** must also deal with the extent to which the learner is actually able to demonstrate this knowledge in a meaningful communicative situation **(performance**). Thus, the ability to use language communicatively consists of both knowledge, or competence, and the capacity for implementing, or executing that competence in appropriate, contextualized communicative language use. So, communicative languge tests are mainly performance based tests, requiring assessment to be carried out when the candidates are engaged in an extended act of communication,either receptive or productive, or both (McNamara, 2000).

Weir (1990:30) lists the following specific features that a test within a communicative paradigm might be expected to exhibit:

- "There would be emphasis on interaction between participants, and the resultant intersubjectivity would determine how the encounter evolves and ends.
- The form and content of the language produced would be, to some extent, unpredictable.
- It would be purposive in the sense of fulfilling some communicative function.
- It would employ domain-relevant texts and **authentic tasks** (section 3.2.1.1).
- Abilities would be assessed within meaningful and developing contexts and a profile of performance on these made available.
- Where deemed appropriate and feasible, there might be an **integration** of the four skills of reading, listening, speaking and writing.
- The appropriateness of language used for the expression of functional meaning would have high importance.
- It would use **direct testing methods** (Section 3.2.1.3)**,** with tasks reflecting realistic discourse processing.
- The assessment of productive abilities would most probably be qualitative rather than quantitative, involving the use of rating scales relating to categories of performance".

3.2.1 Definition of the Key Terms

3.2.1.1 Authenticity

The term 'authenticity' refers to the degree of correspondence between the characteristics of TLU (Target Language Use) tasks and those of the test tasks (Bachman, 1990). Fulcher and Davidson (2007:15) define 'authenticity' as 'the relationship between test task characteristics, and the characteristics of tasks in the real world'. On the ground of authenticity, or approximations to it, integrated tasks demand consideration (Weir, 1990).

3.2.1.2 Integrative Test

An integrative test is one that measures knowledge of a variety of langauge features, modes, or skills simultaneously (Henning, 1987). An example would be dictation, which could be used to measure listening comprehension, spelling, or general langauge proficiency.

3.2.1.3 Direct Test

Direct testing requires the candidate to perform precisely the skill that is supposed to be tested (Hughes, 2003). An *interview* may be thought of as more direct than a *cloze* test for measuring language proficiency. Similarly, a contextualized vocabulary test may be thought more natural and direct than a synonym-matching test. A claim to 'directness' implies a claim for test validity through other concepts such as 'authenticity' (Bachman, 1990).

3.2.1.4 Indirect Test

An indirect test is one that fosters inference about one kind of behaviour or performance through measurement of another related kind of performance (Henning, 1987). Indirect testing measures the abilities that underlie the skills in which the tester is interested (Hughes, 2003). An example would be the measurement of vocabulary use through a test of vocabulary recognition. Indirect techniques are restricted in terms of their perceived validity for test takers and the users of test results (Weir, 1995).

3.3 The Reading Skills

Reading is a complex skill involving the simultaneous practice of a number of different abilities. According to Hughes (2003), the following abilities are required for efficient reading:

- **Skimming** (glancing rapidly through a text to determine its gist)
- **Search Reading** (quickly finding information on a predetermined topic)
- **Scanning** (finding specific words or phrases, figures, percentages; specific items in an index; specific names in a bibliography or a set of references)
- **Careful Reading**

a. i. identifying pronominal reference and discourse markers,

ii. interpreting complex sentences and topic sentences,

iii. outlining logical organization of a text and the development of an argument,

iv. distinguishing general statements from examples,

v. identifying explicitly and implicitly stated central ideas,

vi. recognizing the writer's purpose, and the attitudes and emotions of the writer,

vii. identifying addressee or audience for a text,

viii. identifying what kind of text is involved (e.g. editorial, diary, etc.),

ix. distinguishing fact from opinion, hypothesis from fact, and fact from rumour or hearsay.

b. Making Inferences:

i. inferring the meaning of an unknown word from context;

ii. making propositional informational inferences, answering questions beginning with *who, when, what ;*

iii. making propositional explanatory inferences concerned with motivation, cause, consequence and enablement, answering questions beginning with *why, how;*

iv. making pragmatic inferences.

3.4 The Writing Skills

The ability to write involves at least four component skills (Heaton, 1975:138):

- "**Grammatical skills**: the ability to write correct sentences;
- **Stylistic skills**: the ability to manipulate sentences and use language *effectively*;
- **Mechanical skills**: the ability to use correctly those conventions peculiar to the written language – e.g. punctuation, spelling;
- **Judgement skills**: the ability to write in an appropriate manner for a particular purpose with a particular audience in mind, together with an ability to select, organize and order relevant information".

4. Aims and Objectives of the Course

The central aims of the Higher Secondary language courses are to:

i) 'increase learner motivation by raising awareness that what they are learning is the language of the real world, and is therefore useful to them,

ii) help the learners communicate in a wide range of interesting situations, and

iii) help develop the learners' speaking, listening, reading and writing skills so that they can communicate accurately and appropriately' (*Teacher's Guide*, p. 235).

Thus, the basic objective is to measure the extent to which students have acquired or improved their control of the four major skills for effective communication in real-life situations. In the past, the English language syllabus at the Higher Secondary level had been a selection of prose and poems, and a list of grammar items. Accordingly, tests were based on a few questions from the literary texts, and some discreet-point grammar exercises. So, students were encouraged to memorize the content of the texts rather than to develop their language competence. The existing syllabus is an improvement on the previous one, which is designed with the goal of measuring learners' communicative language ability. The prime objective of this syllabus, as it has been stated in the *Teacher's Guide* (p. 235), is to 'test language as it is used in real life, i.e. **language skills** and **not** memory and **certainly not** the rules of grammar'. With this end in view, samples of communicative language testing have been provided in the *Teacher's Guide* to be treated by the question setters as models.

5. Research Questions

The researchers seek answers to the following research questions:

i) Are the contents of the tests consistent with the stated goal for which the tests are being administered?

ii) How adequately do the test items sample the intended content area?

iii) How well do the tests measure the particular abilities, or constructs, they are purported to measure?

iv) To what extent can the scores derived from these tests be interpreted as determiners of candidates' language abilities?

6. Method

Since the present study aims to respond to research questions of a qualitative nature, data collection and analysis technique from qualitative methodology has been implemented. Qualitative research is fundamentally interpretive, which means that the research outcome is ultimately the product of the researcher's subjective interpretation of the data (Dörnyei, 2007). Qualitative data can be collected from sources like recorded interviews, various types of texts, e. g. field notes, journal and diary entries, documents, etc (Dörnyei, ibid). For the purpose of the present research, three types of documents have been used to collect data:

i) The HSC Test Paper (Question Papers of the HSC Examinations) published in 2015,

ii) The Higher Secondary English Language Syllabus,

iii) The Sample Question Papers in the *Teacher's Guide.*

In this study, an analysis has been made of the test contents (HSC question papers) to see whether the tests contain representative samples of the relevant language skills. In doing the analysis, the effectiveness of the test items used has also been examined in order to determine the extent to which the tests are successful in measuring what they set out to measure. At the same time, the test contents are compared with the syllabus to examine whether the syllabus is fairly

reflected in the contents. Moreover, the sample papers in the *Teacher's Guide* have also been examined to see how far they can help the question setters in setting valid questions.

7. The Test Contents

The Higher Secondary language tests consist of two courses: Paper I & Paper II. All the eight general education boards follow the same question format prescribed by the National Curriculum and Textbook Board (NCTB). Paper I includes 'Seen Comprehension', 'Vocabulary' and 'Guided Writing'. Paper II includes 'Grammar' and 'Composition'. Thus, the tests measure two major skills- reading and writing, and two language elements- vocabulary and grammar. The question types fully adhere to the syllabus specification (Appendix) and the sample question papers in the *Teacher's Guide*.

The reading test is based on short passages taken from the textbook *English for Today for Classes XI and XII,* and a number of tasks to perform which include: i) multiple-choice questions ii) true/false questions iii) filling in gaps (with clues) iv) information transfer (making a list of five points from the ideas contained in the passage, and making a flow-chart v) short-answer questions, vi) filling in gaps (without clues) vii) summarizing the passage in five sentences. According to Heaton (1975), the objective test items, i.e. multiple choice, true/false, gap filling, etc. can never test ability to *communicate* in the target language nor can they evaluate actual performance. Weir (1990) expresses doubt about multiple-choice and true/ false items' validity as measures of reading ability by focusing on the point that candidates answering these items can find the correct response without comprehension of the text by guessing. Moreover, cheating in the examination hall is likely to be easier in answering these items (Hughes, 2003). Weir (ibid) also points out that answering multiple-choice questions is an unreal task, as in real life one is rarely presented with four alternatives from which to make a choice to signal understanding. Among the other items, information transfer tasks resemble real-life activities (Alderson et al., 1995), and are therefore much used in test batteries which try to include authentic tasks. This technique is particularly suitable for testing an understanding of process, classification or narrative sequence and are useful for testing a variety of other text types (Weir, 1990). Items in which the test taker has to provide a short answer are also very common in reading tests. Guessing has a limited effect in such items (Hughes, 2003). According to Weir (1995), this technique is extremely useful for testing reading comprehension. The advantages, he finds, are: i) Answers are not provided for the candidate as in multiple-choice: therefore if a candidate gets the answer right, one is more certain that this has not occurred for reasons other than comprehension of the text. ii) Activities such as inference, recognition of a sequence, comparison and establishing the main idea of a text, require the relating of sentences in a text with other items which may be some distance away in the text. This can be done effectively through short-answer questions where the answer has to be sought rather than being one of those provided. One disadvantage to this technique, however, is that the test taker has to produce language in order to respond (Hughes, 2003). But this disadvantage is not very significant in the Higher Secondary tests, since the required response has been found to be really short. A well-designed summary task is also a very efficient way of testing reading comprehension. Writing summaries may closely replicate many real-life activities (Alderson et al., ibid).

In the reading test, items have been set to measure only the lower-order skills of reading, such as– skimming, scanning, search reading, etc.; higher-order sub-skills are not included. As a result, this test would not be sufficient to justify using it as an achievement test for a reading course. Moreover, a number of problems have been identified in the selection of comprehension passages. Hughes's (2003) suggestions for the testers are to avoid texts made up of information that may be part of candidates' general knowledge and to avoid those that candidates have already read. But it has been found that many of the questions in this test can be answered from general knowledge without reading the text. The reading passages in the question papers have been scrutinized, and it is found that the same passages have been used in most of the question papers. Although the textbook contains twenty-four units, the passages are taken from only six or seven units (unit-3, 6, 9, 10, 13, 21, etc). A few selected passages have been repeated so many times that candidates do not need to have any reading ability for taking this test, since they can easily respond by memorizing answers. Though it is clearly recommended in the syllabus (Appendix) that the test should not encourage memorization, the question setters have turned the test fully into a test of memory.

Writing skills are tested in both paper I and II. 'Guided Writing' in Paper I includes tasks, such as: i) Matching- Candidates are asked to make six meaningful sentences by matching some phrases in a substitution table, ii) Ordering- Examinees are required to rearrange 14 "scrambled" sentences into a coherent paragraph. iii) Paragraph- Candidates are asked to write a paragraph of about 200 words, answering 6 or 7 questions. Here, the first two items are good for testing specific micro-skills of writing, such as organizing and ordering skills. But such tasks have little communicative value, since they do not require production. A disadvantage with the matching item is that once all the sentences except the last one have been accurately formed, the last pair is correct by default (Alderson et al., 1995). Paragraph organization item is also less effective as predictors of general writing skill, since it is extremely difficult to find or compose paragraphs which can be reordered in just one acceptable way (Harris, 1969). Moreover, these two items are used as indirect measures of writing ability. As far as achievement tests are concerned, it is preferable, according to Hughes (2003) to rely principally on direct testing. Moreover, as we have seen, tests of communicative language ability should be as direct as possible, and the tasks candidates have to perform should involve realistic discourse processing.

The most direct way of measuring students' writing ability is to have them write (Hughes, ibid). The composition test in Paper II attempts to 'directly' measure the construct of interest by asking candidates to write paragraph, short composition, application, dialogue and report. But the tasks do not provide any context, or background information, and thus lack authenticity. Weir (1990) considers this type of free, open-ended writings to be invalid tests of writing ability. According to him, the writing component of any test should concentrate on controlled writing tasks where features of

audience, medium, setting and purpose can be more clearly specified. Heaton (1975: 128) argues: "how often in real-life situations does a person begins to write when he has nothing to write, no purpose in writing and no audience in mind"? This view is fully supported by Douglas (2010) who says, language is never used in a vacuum; we don't simply speak, write, read, or listen. We always do so for a purpose, related to the context, the situation we are in. According to Douglas (ibid), if the test purpose is to make inferences about a learner's language ability in some communicative context, then the test should provide relevant contextual information. It is important, he believes, to establish a context for language use in our tests to avoid the test takers imagining their own and thus making our interpretations of their performance potentially wrong.

The setting of the writing tasks, as we can see, is highly problematic. To improve the effectiveness of a composition test, Harris's (1969) suggestion is to set tasks that are within the reach of all, since the purpose of testing is to measure only writing ability, and nothing else. But it is found that some topics are set which are discipline-specific, e.g. 'Greenhouse Effect', 'Deforestation and its Devastating Effects on the Environment', Environment Pollution, The Dangers of Drug Addiction, etc. The ability to write on some others, e. g. 'A Victory Day Celebrated in Tangail', 'A Book Fair Held in Bangla Academy Premises', 'A Cultural Week Observed in Nilgonj Govt. Mohila College', etc. depend on the candidates' background, or cultural knowledge. A few others favour candidates who have wide general knowledge, e.g. 'World Cup Cricket in Bangladesh', 'War of Independence', 'Price-hike in Bangladesh', 'SIDR', 'Rural Development', 'Unemployment Problem', etc. Moreover, the topics that are given for writing are all available in the guide-book popularly known as *H.S.C English Grammar and Composition* which includes so-called model paragraphs, essays, applications, etc. on a limited number of topics. So, students who are good at memorizing texts can easily obtain good marks in this test. The topics are also highly repetitive.

The following tables (Table-A & Table-B) show some frequently used topics of paragraphs and dialogues found in the question papers of HSC Final Examinations of different educational boards within the last four years (2012-2015).

Table **A**

Topics for Writing Paragraph	**Frequency of Repetition**
Your Country	7
Dowry System	4
A Book Fair	4
Female Education	3
Eve-teasing	3
Merits and Demerits of Satellite TV Channels	3

Table B

Issues of Writing Dialogues	**Frequency of Repetition**
Between two friends on the choice of a career	6
Between two friends stating the causes of students' failure in English	5
Between two friends about the uses and abuses of mobile phone	4
Between two friends about the dangers of smoking	4

In the testing of writing, skills involving the use of judgement are of far greater importance than those concerned with the correct use of language, or the effective use of language (Heaton, 1975). Here, we have found that candidates' grammatical skill is highly emphasized. Nearly half of the total marks (40 out of 100) have been allocated for grammar, and a variety of test items have been used, e.g. right form of verbs, use of prepositions, use of articles, filling in gaps with suitable linking words, phrases and idioms, rewriting sentences in indirect speech, transforming sentences, making tag questions, completing sentences, etc. While grammar contributes to communicative skills, it is rarely to be regarded as end in itself (Hughes 2003). Too much concentration on the testing of this language element may create a harmful effect that undermines the achievement of the objectives of teaching and learning where these are communicative in nature (Hughes, ibid).

The vocabulary test consists of only one test technique, i.e. cloze test. Two cloze passages are given, one with clues and one without clues, and the testees are asked to fill in gaps in the given passages. There are many types of items that can be used to test vocabulary, e. g. sets (associated words), matching items, word formation test items, items involving synonyms, rearrangement items, definitions, completion items, etc (Heaton, 1975). So, what becomes clear is that with so many potential tasks and with only one item, the test's content validity is inevitably brought into question. Moreover, this test has very little to do with testing vocabulary, because the test words are drawn directly from the set textbook

without any change. So, the candidates do not need to look for suitable words for the gaps as they have already encountered the passages in the textbook. Thus, the test may be helpful for developing learners' passive vocabulary to some extent, but the learners cannot be expected to be able to apply or use those words in the right context when speaking or writing.

A few integrated tasks have been found in the reading and grammar tests. The open-ended questions and summary writing, for example, are primarily concerned with developing reading skills, but they also help develop writing skill. Similarly, completion items in the grammar test integrate grammar and writing skill.

8. Answers to the Research Questions

The findings of the study in relation to the research questions are as follows:

Language tests at the Higher Secondary Level in Bangladesh are highly selective and limited in the objectives they measure. The analysis of the test contents clearly shows that the demands made on the candidates taking these tests are not appropriate and in accordance with the stated aims of the course. As we have seen, the test items and test techniques do not fit well with the objectives. The curriculum claims to be a competency-based one whereas the associated assessment procedures are designed to assess bookish knowledge of the candidates. The curriculum stresses the need for students to learn to communicate in English rather than just to master the structure of the language (National Curriculum). But the tests are not appropriate at all for measuring students' attainment in terms of their communicative ability. The basic aim of testing is to assess the ability in the context of simulations of real-world tasks in realistic contexts. But the tests do not reflect the principles underlying the communicative approach. Thus, there is a great inconsistency between the test purpose and the testing system. Presently, though there has been an attempt to contextualize the grammar and vocabulary in the form of cloze passages, the writing components are totally devoid of context. The integrated approach has not been adopted in any effective way as it would have been if the tests were constructed within a model of what constitutes communicative demand. Integrative tests like dictation, translation, oral interviews and conversation (as suggested by Oller, 1976), which could have involved a simultaneous testing of the testees' multiple types of competences from various perspectives, are all absent in the test format. A quick glance at the syllabus (Appendix) reveals that the syllabus itself is not consistent with course objectives. Candidates' aural/oral skills have completely been ignored in it. Even the sample papers in the *Teacher's Guide* themselves are not in line with the communicative testing approaches. So, it is obvious that successful performance on the Higher Secondary tests do not indicate successful achievement of the objectives.

The language tests now in use at the Higher Secondary level do not include representative samples of all the language skills with which they are meant to be concerned. In order to have content validity, the tests must cover an adequate and representative section of those areas and skills they are desired to test. But our study has found the test contents to be in favour of the less important items to the exclusion of the most important ones. The syllabus specification seems to be biased towards the kinds of items which are easiest to write or towards the test material which happens to be available. Content validation of the tests might confirm that all the four communication skills were well represented in the tests. But only two out of the four skills are tested, and even the tests of the two skills covered do not reflect all the areas of assessment in suitable proportions. The reading component includes testing of some lower level sub-skills; higher level skills of reading have been neglected. Test items in the writing section focus less on guided but more on a free form of writing. Thus, the concentration is on testing only those areas which most easily lend themselves to testing. Testing of vocabulary has been confined to the same type of task (cloze procedure) for many years. So, it is clear that these contents cannot form a satisfactory basis for the inferences to be made from test performance.

The contents of the Higher Secondary tests do not adequately reflect the breadth or depth of the *construct* as defined for the purpose of these tests. We know that construct validity assumes the existence of certain learning theories or constructs underlying the acquisition of abilities and skills. So, the greater the relationship which can be demonstrated between a test of communicative competence in a language and the theory of communicative competence, the greater the construct validity of the test. Our analysis shows that the items used do not reflect the essential aspects of the theory on which the tests are based. The tests are found to have both construct under-representation and construct irrelevant variance characteristics. The tests have failed to capture the complexity of the communicative demands of the criterion, i.e. the 'real-world' performance, in which the tests are trying to predict ability to succeed. As a result, the tests require too little of the candidates. Only one aspect of communicative competence, i.e. grammatical competence has been focused in these tests. In this case, defining the construct to include only one area of language knowledge is inappropriately narrow, since the construct involved in the TLU (Target Language Use) domain- ability to perform academic writing tasks- involves other areas of language knowledge. Our theoretical framework suggests that the tests should emphasize performance. Our analysis reveals that the tests do not give any indication of candidates' skills in performing in actual communicative situations, since most of the tasks are inauthentic, and no sample of speech is elicited from the candidates. The reading test is an inadequate measure of reading ability. The tasks are unable to involve candidates in providing evidence of successful reading. The theory of reading states, as we have seen, that there are many different constructs involved in reading and that the constructs are different from one another. But only a few of them are tested. The tests also have construct irrelevant difficulty in score interpretation. They introduce factors which are irrelevant to the aspect of the ability being measured and which may cause changes in test scores. As we can see, performance on the reading test may give a quite inaccurate picture of the candidates' ability. Scores on some of the items used in this test may be invalidly high because of the ease of guessing correct answers when the answer is not known. Some items can be answered from general knowledge whereas some others facilitate cheating in the

examination hall. Thus, the scores can be unduly affected by many factors other than the ability being tested. The major drawback of these tests is that they are all memory tests. From the high frequency of repetition of questions, it would be quite easy for the candidates to predict the questions beforehand resulting memorization. Thus, tests which are meant to measure candidates' communicative ability are actually measuring such abilities as guessing, memorisation, general knowledge, etc.

As already stated, in examining validity, we must be concerned with the appropriateness and usefulness of the test score for the given purpose. In order to justify a particular score interpretation, we need to provide evidence that the test score reflects the area(s) of language ability we want to measure, and nothing else. From the above discussion, it is clear to us that the scores obtained by the test takers from the HSC language tests do not reflect the degree of presence or absence of the construct concerned, since the tests do not measure the abilities they are supposed to measure. Scores derived from tests measuring candidates' capacity to memorize text can not be appropriate for determining candidates' language abilities. So, the scores are not meaningful, and cannot, therefore, provide the basis for valid interpretation or use.

9. Recommendations

- A set of specifications should be developed involving assessment of all the major language skills. The 'content' section in the specifications should include tasks in each skill area, which directly attempt to *simulate* appropriate real-life operations, with contextually appropriate conditions and which can be assessed by relevant target situation criteria. The syllabus should be designed in accordance with the specifications and should be fairly reflected in the test contents.
- The reading test should be constructed in such way that it tests more than a superficial understanding of the text and requires the candidates to digest and interpret what they have read. Each test question should sample one or more of the reading abilities listed in **Section 3.3**. Test writers should try to achieve a balance so that one or two skills are not over-tested at the expense of the others.
- Tasks which are unable to elicit valid samples of writing should be excluded from the test format. More direct extended writing tasks of various types should be constructed.
- For the vocabulary test, items should be chosen widely from the whole area of content.
- Test of grammatical ability should not be given too much prominence in relation to tests of skills, the development of which constitutes the main objective of the course. Marks allocation should be revised with less emphasis on grammar and more on language skill test items.
- Items should be sampled unpredictably so that the candidates do not get scope to write memorized answers.
- Samples of valid question paper should be provided in the *Teacher's Guide*.
- Language teachers should be given training on how to construct valid test items.

10. Conclusion

This study has identified a number of sources of invalidity of the language tests used at the Higher Secondary level in Bangladesh. The findings indicate that successful performance on these tests is not enough to ensure a high degree of language ability. Because the tests are neither valid tests of stipulated objectives nor are they valid measures of communicative competence. The real requirements of the criterion are not fully represented in these tests. The constructs tested are not enough and appropriate to ensure that the scores obtained by the candidates represent the abilities that the candidates are supposed to have. The items used do not at all have anything to do with any of the language skills that the learners need to acquire or master. To be able to successfully respond to these items, what the candidates need is only to memorize the answers to a few selected questions and to reproduce them in the examination hall. So, the scores are not relevant to what is being tested and can be affected by many abilities other than the ones that are genuinely required. To conclude, it can be said that unless we can demonstrate that the inferences we make on the basis of these tests are valid, we have no justification for using test scores as the indicators of candidates' abilities. Here, the authority concerned has a great responsibility to actively promote the responsible use of tests and the appropriate interpretation of test performance, and should, therefore, take immediate steps to implement the recommendations mentioned above.

References

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

American Psychological Association. (1985). *Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association.

Anastasi, A. (1982). *Psychological Testing*. London: Macmillan.

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

Canale, M. (1983). *Language and Communication*. In J. C. Richards and R. W. Schmidt (eds.). London: Longman.

Canale, M. and Swain, M. (1980). 'Theoretical Basis of Communicative Approaches to Second Language Teaching and Testing', *Applied Linguistics,* 1, 1-47.

Dörnyei, Z. (2007). *Research Methods in Applied Linguistics.* Oxford: Oxford University Press.

Douglas, D. (2010). *Understanding Language Testing.* Routledge. New York.

Ebel, R. L. and Frisbie, D. A. (1991). *Essentials of Educational Measurement*. 5th edition. Englewood Cliffs, NJ: Prentice-Hall.

Fulcher, G. (2010). *Practical Language Testing.* London, UK: Hodder Education.

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

Gifford, B. (1989). *Test Policy and Test Performance: Education, Language, and Culture*. University of California, Berkeley.

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

Harrison, A. (1983). *A Language Testing Handbook.* Hong Kong: Macmillan Publishers Ltd.

Heaton, G.B. (1975). *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.

Hymes, D. (1972). 'On Communicative Competence'. In Pride J. B. J. and J. Holmes (eds.), *Sociolinguistics: Selected Readings*. Harmondsworth: Penguin, 269-93.

Islam, J. et. al. (2002). *Teacher's Guide.* ELTIP in Association with the British Council.

Kabir, M. H. (2009). 'How Validity is Ensured in Our Language Test: A Case Study', *IIUC STUDIES*, 5, 37-52.

Kane, M. T. (2006). 'Validation'. In Brennan, R. L. (ed.), *Educational Measurement.* 4th edition. New York: American Council on Education/Praeger, 17-64.

Kerlinger, F. N. (1973). *Foundations of Behavioural Research.* New York: Holt, Rinehart and Winston.

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

Morrow, K. E. (1979). 'Communicative Language Testing: Revolution or Evolution?' In Brumfit, C. J. and Johnson K (eds.), Oxford: Oxford University Press.

National Curriculum- English: Classes XI & XII. Dhaka: NCTB. pp. 134-153.

Oller, J. W. (1976). 'Language Testing.' In Wardhaugh, R. and Brown, H. D. (eds.), *A Survey of Applied Linguistics.* Ann Arbor: University of Michigan Press.

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

Appendix

Syllabus for Classes XI & XII (Prescribed by the National Curriculum & Textbook Board)

ENGLISH (PAPER I)

1. Seen Comprehension

There will be a seen comprehension passage followed by a choice of questions. The questions should be equally divided between objective and more free/open. Comprehension question types should include the following.

a. Objective: (i) Multiple choice, (ii) True/False, (iii) Filling in gaps with clues, (iv) Information transfer, (v) Making sentences from substitution table (s) (vi) Matching phrases/pictures, etc.

b. **More free**: (vii) Open-ended, (viii) Filling in gaps without clues, (ix) Summarizing, (x) Making notes, and (xi) Re-writing in a different form.

All the questions should test the students' ability to understand the passage as a whole, rather than their ability to copy section from it. Although the seen comprehension passage will be taken from a set textbook, it will not encourage memorization, because (i) the passage will not be reproduced in the question paper, and (ii) the questions will not be taken from the textbook, but rather, will be new.

2. Vocabulary

There will be question on vocabulary contextualized in the form of cloze passage with clues, and cloze passage without clues. In order to provide more communicative contexts, the topics should be related to those already encountered by the students in seen and unseen comprehension.

3. Guided Writing

There will be a number of writing tasks. The following types of exercises should be given:

(i) Producing sentences from substitution tables,

(ii) Reordering sentences, and

(iii) Answering questions in a paragraph.

Distribution of Marks

a.	**Seen Comprehension**	**40 Marks**
	Objective questions	20
	More free/open questions	20
b.	**Vocabulary**	**20 Marks**
	Cloze test with clues	10
	Cloze test without clues	10
c.	**Guided Writing**	**40 Marks**
	Total Marks= 100	

ENGLISH (PAPER II)

A. Grammar (Note: Grammar items introduced previously may be used if needed)

1. Modifiers
2. Questions with noun, adverb and adjective (e.g. Do you know what time the shop closes?)
3. Use of adjectives and adverbs (e.g. Adjectives without comparative & superlative forms, and using words both as adjectives and adverbs)
4. Word formation (suffix— prefix)
5. Sentence structure
6. Use of modals
7. Subject-verb agreement
8. Use of direct & indirect speech
9. Transformation of sentences
10. Use of tenses (e.g. used to, to be + used to, etc.)
11. Linking words
12. Appropriate words
13. Use of articles
14. Idioms and phrases
15. Emphatic statements (e.g. Do come here.)
16. Use of infinitives and participles

B. Composition (Note: Focus should be on the practice of the grammar points introduced in section A above, as far as possible):

1. Writing instructions
2. Writing summaries
3. Writing arguments logically and clearly
4. Writing composition about ceremonies, festivals, events, travel experiences, topics of public interest, environment, etc
5. Writing composition using charts, pictures, graphs, etc
6. Completing a story/passage
7. Writing a dialogue on a given situation
8. Writing informal/formal letters (including job applications, filling in forms, CV and e-mail)
9. Writing composition on imaginary situation
10. Writing reports

Distribution of Marks

A Grammar : 40	**B. Composition: 60**	
Each type of test item carries 5 marks.	Paragraph/Report	10
Any 8 items out of 9 will be tested in an	Short composition	15
examination.	Completing a story	15
	Writing a summary/dialogue	10
	Formal letter	10