NAEM JOURNAL
ISSN 1997-4248, eISSN 2959-4561
Volume 17 Issue 02
Serial 34
https://Doi.org/10.59629/NAEM 20213202.1601

Assessing the assessment system in Bangladesh: An investigation into English Language Assessment of Secondary School Certificate Examination (SSC)

Dr. Md. Mostafizur Rahman[1]
Dr. Md. Habibullah[2]
Md. Masud Rana[3]

Abstract

The study aimed at reconsidering the challenges of English language assessment system of the Secondary School Certificate (SSC) examination in Bangladesh. Attempts have been made to examine strengths and weaknesses each test item of SSC English first paper. The English first paper comprised of twelve items is designed to achieve ends of reading and writing skills of students. A noticeable gap has been found between skill getting and skill using of students who have already passed in the SSC public examination. Considering the issue the research team focused on the test construction and performance of each item. Moreover, the study laid importance on strategies of item designing without which flaws remain attached with the item for which a particular item fails to carry out its function individually and in a group. To meet this end the research project evaluated scores of the students who are examinees of the SSC-2022. In this regard, primary data was collected and analyzed quantitatively from the test administration to aforementioned examinees and thematic analyzed was done qualitatively based on questionnaires served to the question setters and moderators. In addition, the research team carried out an investigation into secondary data sources related to existing curriculum, syllabus and instruction manual of question setters and moderators. The study, moreover, took into consideration the prevailing assessment system which projected a visible dearth of reliability. Hence, the central focus of this study was to find out the challenging areas which are difficulty level, discrimination index, validity and reliability associated with test construction. The study followed the methods of empirical research which considers the process of observation, exploration, analysis and interpretation. Both the qualitative and quantitative methods were applied to explore the objectives of this study. Finally, based on the findings of this study the researchers offered some recommendations on the enhancement of validity and reliability of SSC level quality test construction and particularly on answer scripts assessment so that the test score matches with the competency of the learners.

Key words: Assessment, Evaluation, Validity, Reliability, Difficulty, Discrimination Index, Interpretation and Competency

[1]Senior lecturer, Rajshahi Gov.Model school and college, Rajshahi, bangladesh.
[2]Professor, Dept of political science, Varendra University
[3]Teacher Trainer, NAEM

1.0 Introduction

The English language test construction, administration and evaluation are an uphill task in case of high stake public examination. In the above process, each plays a crucial role to help others to make the SSC examination a successful one. Moreover, with respect to designing question paper, the question setters and moderators play a vital role in this ragard. The language teachers are supposed to prepare their students by facilitating four skills together but in most cases they lay emphasis on reading and writing skills. Unfortunately listening and speaking are vehemently ignored by them like Cinderella. Thus the mission to learn language in an integrated manner remains unsuccessful. The students with half and erring knowledge of language face the public examination. . Hence, teaching approaches, methods and techniques claim a meticulous investigation in this process of research. Recently critics have the issue that in teaching-learning enterprise a large number of teachers are testing what they are not teaching. What is more menacing is that texts remain unread, on the other side guidebooks and help books flourish preparing predictable and expected questions. However, the test administration process has not yet been reached at the minimum level of standard since either school environment or public examination system has not given a favorable atmosphere to take the test of four skills together. The education administration and board authority have not been able to ensure the matter of holding the examination in accordance with the principles of Communicative system of Teaching and Testing. This is how a vital issue appears that no attempts are made to bridge the gap between linguistic competence and communicative competence. Even neither from the part of language teachers nor from education administration develops a smooth transition between skill getting and skill using. In an educational set up, in the large-scale, public examination evaluation plays a major role in the award of marks, grades and degrees. In process of evaluation, learners know their potentials, enable them to learn more about their desired topics and acclimatize creative power when they encounter challenges. This research explored depth, breadth and dimensions all sorts of summative and formative evaluation so that language teachers, teachers and teacher-educators might find the path to improve their route of evaluation of the public and school-based examination. Moreover, attempts were be made to investigate existing principles and duties of evaluators and head examiners so that they might also find innovative strategies in their ways of evaluation. More importantly in the SSC level test analysis and assessment pays a crucial issue and it is one of the most important areas in teaching learning process. It is the process of defining, selecting, designing, collecting, analyzing, interpreting, and using information to increase students' learning and development. If the assessment process is sound, it can show an authentic picture of the learners' actual strength and abilities. However, in the field of education test as a part of evaluation process plays a vital role identifying trend and potentials of learners. This research took into account the varied number of tests prevalent across the world and their application in Bangladeshi context. Besides this study would focus on different types of language test and a special attention was be given on SSC level language test items designed for examination paper to explore its individual strengths and limitations. Hence to find out the reasons of this anomaly plagued in the testing and evaluation systems is the demand of the time. When the total process of assessment is questioned, all these related items come under the purview of

language testing and assessment. Therefore, it is obvious to carry out research to bring forth the existing scenario of assessment for the sake of an authentic teaching learning atmosphere of Bangladesh in aforementioned level.

Objectives of the Study W

First, find out to what extent the test items of SSC level English language question papers employed in the public examination are following the areas of content, construct and face validity.

- Second, to find out the reliability level in terms of evaluation procedures and score assessment.
- Finally, to analyze and examine the policy documents of curriculum and syllabus, question setters, evaluators and moderators.

Review of Literature

Validity and reliability are essentially associated with test construction and assessment. Similarly difficulty level and discrimination index are two vital in the field language testing and assessment. This study addressed above-mentioned four academic issues to examine the strengths, qualities and weaknesses of SSC level language test. In this respect, Aggarwal (1986) is of the view that "another step which leads the calculation of item difficulty and item discrimination of a test is item selection based upon the judgment of competent persons as to the suitability of the item for the purposes of the test." However, in course of the analysis test items employed in the SSC level English language test paper, which we consider for a high stake achievement test through which students are issued with certificate based on the ability in two skills-reading and writing. A valid test is a tool, on which we depend much, to achieve sound judgment. Therefore, some of the views by the experts are as follows - "A test is valid when it measures what it claims to measure" (Garret, 1964: 30). Again, "a test is valid when it measures what it ought to measure"(Ebel, 1972: 463).Besides, validity of a test deserves special attention. Validity can be defined as the suitability or the appropriateness of a test for its purpose. The validity of a test has different aspects. The major aspects are content validity, construct validity, ethical validity, consequential validity, concurrent validity, predictive validity and face validity. Moreover, the researcher would meticulously endeavor to find and measure all aspects of validity in the process of the present study.

Furthermore, reliability and standardized tests matter a lot in a high stake achievement test, which claim test items to be reliable. There are many high stake tests which are purely system oriented and which are not owned by a single teacher or a paper-setter. They are created by a group of people and not made to be taken only by a single group of students. In addition standardized tests, according to Anderson (1975:384), attempts are made to provide for a measurement of individual differences in as unambiguous ways, as possible. Thus, the process of standardization allows all aspects of testing - construction, administration, scoring, reporting, and interpretation of results. Hughes (2003) and Hyland (2006) have harped on the need for training of assessors to enhance the reliability. Hence reliability is no longer regarded as a separate characteristic of test; it is treated as an aspect of validity itself.

A qualitative study was conducted by Rubina Khan on English Language assessment on English subject in Bangladesh where she has shown the strength and weaknesses of SSC level question items and she also projected the washback effect of the ssc level public examination (Khan, R. 2010). It is worth mentioning, that wash back relates to testing on teaching and learning; and it heavily influences teachers, curricula and materials. Hamp Lyons (2000: 586) rightly points out that, "impact deals with the wider influence of tests, their consequences beyond the classroom." Finally, for smooth and sound teaching, learning and testing, there needs to be a clear framework for test modeling so that content, structure, format, scoring procedures are explicit. In this regard, Bachman (1990) gives a comprehensive checklist for test specifications. In process of this study the researchers would take into account all the checklists to examine the prevailing test and assessment procedures to find out the gap between the existing test types and the mentioned areas. Hence there is a rationale for Checking reliability and validity together in this Study. Reliability and validity are the most fundamental issues in any measurement. Any statistical measurement can bear proper weightage with the appropriate application of the both. Each is, hence, an integral part of other. Validity remains insignificant without the proper intervention of reliability and vise-versa.

3. Methodology

The study followed the methods of imperical research which considers the process of observation, exploration , analysis and interpretation. both the qualitative and quantative methods were applied.

Sampling

The researchers have selected 10 schools from education board. The number of schools has been chosen both from urban and rural area. All the teachers and students of these 10 schools were taken as the sample of this study. However, research will be conducted on at least 200 students, 100 students from each board. Besides, 10 practicing teachers were selected from the aforementioned boards, 5 teachers from each board. Moreover, 10 question-setters, 10 moderators and 10 evaluators /examiners were contacted for soliciting information out of two education boards, 5 from each category from each board. However, these schools were selected as random sampling process. But it is worth mentioning that research team carefully tried to collect data almost equal number both from MPO and Government schools to make the information authentic. However, since the number of MPO schools is more than that of government ones, more data came from MPO schools.

Source of data and method of data collection

Primacy Sources:

The team endeavord to administer question papers to the SSC level students who have already prepared themselves to appear in the SSC examination. Attempts was made to design questionnaires to assess their abilities in English as foreign language. Besides the researcher designd separate questionnaires for question-setter, evaluators and moderators who were inevitably associated with test construction, administration and evaluation process.

Secondary Sources:

The research team investigated documents available in the two boards regarding the policies and principles of question-setter, moderator and evaluators. The process helped them to collect all the written materials relating to duties, responsibilities, principles and techniques of question formation, moderation and evaluation.

Methods of data analysis and presentation:

The validity and reliability are the main issues which are to be elicited and examined in order to show the prevailing state of the question items. The statistical procedure is one way to demonstrate and define the level of validity and reliability. Moreover, in order to extract level of validity and reliability, showcasing the data characteristics are the pre-requisite issues which indicate the symmetric and asymmetric level of data. To meet these crucial ends the statistical calculation such as mean, standard deviation and standard error level were measured as a vital issue of this study. Furthermore, to demonstrate different levels of validity and varied shades of reliability the statistical procedures such as correlation co-efficient and Cronbach's Alpha tools measurement are essential. More importantly, the SPSS software program was used to reveal the results of the content construct and face validity of existing tests. In the same way, the software also helped to interpret status of reliability based on the assessment of students' scores.

Data analysis and Findings

Item Analysis in Terms of Reliability:

1. Chronbach's Alpha is a way to measure the internal consistency of a questionnaire or survey.

2. Cronbach's Alpha ranges between 0 and 1, with higher values indicating that the survey or questionnaire is more reliable.

An examination was conducted to elicit the information about the level of reliability of the question items used to know reading and writing skills of English language test of the SSC level.

A total number 201 student participated in the English first paper examination process. The number of student has answered to 12 items of English first paper question.

The researchers examined internal consistency based on the following table which indicates the Cronbach's Alpha and Level of Consistency amongst the items against the total. The SPSS calculation by using Cronbach's Alpha theoretical tool has been able to show reason if any item needs to be removed or retained to ascertain internal consistency. The following table containing Cronbach's Alpha values attempts to see what would happen to reliability analysis if one or two items are deleted or corrected from the scale and also would help identify any items which may not have the performing ability to belong to the list of the questions.

The reliability of the 12 following items have been checked through SPSS software to understand level of reliability by means of Cronbach's Alpha internal consistency parameters.

The following table describes how different values of Cronbach's Alpha are usually interpreted:

Cronbach's Alpha	**Internal consistency**	
$0.9 \leq \alpha$	Excellent	
$0.8 \leq \alpha < 0.9$	Good	
$0.7 \leq \alpha < 0.8$	Acceptable	
$0.6 \leq \alpha < 0.7$	Questionable	
$0.5 \leq \alpha < 0.6$	Poor	
$\alpha < 0.5$	Unacceptable	

Table-11(a) : Cronbach's Alpha range

The researchers selected a question paper (English 1st paper) which was administered in all Education Boards in Bangladesh. A reliability analysis has been carried out through measurement of Cronbach Alpha by using SPSS software on the 12 items of this question. The names of these items are as follows:

q1 (MCQ), q2 (Short question), q3 (gap fill), q4 (complete the information) and q5 (Summary writing), q6 (Matching the sentences), q7 (re-arrange sentences and make story), q8 (writing paragraph), q9 (completing story), q10 (Describing graph), q11 (letter writing) and q12 (Dialogue writing).

The reliability can be measured through the scores of Cronbach's alpha, inter-item correlation, item-total correlations, and Cronbach's alpha if item deleted. Cronbach's alpha will explain how closely a set of test items function in a group.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
0.798	0.819	12

Table-11(b): Cronbach's Alpha related to reliability of test item

Table-11(c): Total item statistics for measuring Reliability

Interpretation:

The above table projects that the number of items is 12 and the value of Cronbach's α is 0.798 which indicates that the consistency of the scores is remained in the acceptable range. Hence the items belonging to both reading and writing parts are acceptable in term of relevance.

The following above table demonstrates the Cronbach's Alpha values of both columns- Cronbach's Alpha if Item Deleted and Corrected Item-Total Correlation respectively.

In case of 12 items, the Cronbach's Alpha values have reached α = 0.798 and the most of the items have hovered around the same Cronbach's Alpha α = 0.798 except a few items having a little less than 0.798. Hence, according to the expected reliability range item in the question paper bears high reliability. Therefore, it is presumed that most of the items can be taken good enough and can be regarded worthy of retention. However, the Cronbach's Alpha α of q7 (if deleted) is higher than 0.798, the items can be regarded as unreasonable items and they can be deleted. On the other hand, the items belonging to the group considered as good items and can be taken as functional ones for the purposes. As a consequence one out of 12 items, in the test paper, the value 0.801 is higher than the rest of the items, which points out that there are some problems in the q7, re-arrange sentences and make story. It is well assumed that the q7 is too easy or too difficult to some extent. Hence, the Q7 claims structural and logical improvement to correspond to the entire test paper.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q01	61.7761	196.720	0.479	0.531	0.787
Q02	61.2960	161.861	0.659	0.548	0.758
Q03	65.3731	196.585	0.319	0.439	0.794
Q04	63.7164	197.659	0.473	0.399	0.788
Q05	61.0697	196.418	0.351	0.222	0.792
Q06	63.7861	191.739	0.494	0.578	0.783
Q07	62.0846	193.308	0.255	0.484	0.801
Q08	61.4677	173.605	0.634	0.493	0.765
Q09	61.9925	163.099	0.660	0.513	0.759
Q10	62.3333	167.771	0.499	0.425	0.779
Q11	62.3607	174.593	0.381	0.446	0.795
Q12	62.0572	182.655	0.350	0.489	0.795

Validity Measurement:

The research addressed the issue of item analysis from the perspective of validity on the SSC English First Paper of 12 items which are q1 (MCQ), q2 (Short question), q3 (gap fill), q4 (complete the information) and q5 (Summary writing), q6 (Matching the sentences), q7 (re-arrange sentences and make story), q8 (writing paragraph), q9 (completing story), q10 (Describing graph), q11 (letter writing) and q12 (Dialogue writing).

Item Analysis in Terms of Validity:

The present English first paper test deserves to be checked in terms of the construct validity. Besides, it is imperative to look into a measure of the construct and how well different items correlate with each other. With the end in mind the researchers put forward the three criteria with a view to projecting the level of validity of the each individual test with its relation to the others. Besides, the researchers would like to look into how well a test functions in a total test procedure in a test paper. Moreover, the aim of this study is to see the construct validity which investigates attribute or task lies in the given tests to examine skills (reading & writing) of which the students have got and applied in the test procedures. To find construct validity in the test procedure is one of the crucial measures in the validation process of a test to determine the level of effectiveness of this tool of evaluation.

The research in terms of validity lays emphasis on the internal correlation. Besides, the study should look at the 3 types of internal correlation which can be listed as follows.

Criterion1. Firstly, the correlation between different components should be low (0.3-0.5). Criterion 2. Secondly, the correlation between two tasks in a testing component should be high (at least 0.5-0.7).

Criterion 3. Thirdly, each component should have a high correlation coefficient with the total score (above 0.7)

Exactly –1	A perfect downhill (negative)
–0.70	A strong downhill (negative)
–0.50	A moderate downhill (negative) average
–0.30	A weak downhill (negative)
0.	No linear relationship
+0.30	A weak uphill (positive)
+0.50	A moderate uphill (positive) average
+0.70	A strong uphill (positive
Exactly +1	A perfect uphill (positive)

Table-12 (a): Pearson Correlation

		Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08	Q09	Q10	Q11	Q12	TOTAL
Q01 MCQ	Pearson Correlation	1	.341**	.541**	.543**	.268**	.592**	.440**	.305**	.259**	.224**	0.010	0.019	.540**
	Sig. (2-tailed)		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.886	0.786	0.000
	N	201	201	201	201	201	201	201	201	201	201	201	201	201
Q02 Short Q	Pearson Correlation	.341**	1	.209**	.327**	.286**	.453**	.220**	.622**	.553**	.337**	.281**	.423**	.756**
	Sig. (2-tailed)	0.000		0.003	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000
	N	201	201	201	201	201	201	201	201	201	201	201	201	201
Q03 Read the text & Fill the gap	Pearson Correlation	.541**	.209**	1	.396**	.178*	.467**	.347**	.234**	.218**	.258**	-0.068	-.195**	.417**
	Sig. (2-tailed)	0.000	0.003		0.000	0.012	0.000	0.000	0.001	0.002	0.000	0.337	0.006	0.000
	N	201	201	201	201	201	201	201	201	201	201	201	201	201
Q04 Complete the table from the infor of the passage	Pearson Correlation	.543**	.327**	.396**	1	.206**	.433**	.181**	.349**	.268**	.279**	.202**	0.107	.532**
	Sig. (2-tailed)	0.000	0.000	0.000		0.003	0.000	0.010	0.000	0.000	0.000	0.004	0.130	0.000
	N	201	201	201	201	201	201	201	201	201	201	201	201	201
Q05 Write Summary	Pearson Correlation	.268**	.286**	.178*	.206**	1	.227**	.381**	.261**	.183**	.237**	0.088	0.061	.441**
	Sig. (2-tailed)	0.000	0.000	0.012	0.003		0.001	0.000	0.000	0.009	0.001	0.215	0.393	0.000
	N	201	201	201	201	201	201	201	201	201	201	201	201	201
Q06 Match Sentence from the Columns	Pearson Correlation	.592**	.453**	.467**	.433**	.227**	1	.543**	.351**	.274**	0.087	0.043	0.113	.568**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.001		0.000	0.000	0.000	0.217	0.545	0.111	0.000
	N	201	201	201	201	201	201	201	201	201	201	201	201	201
Q07 Rearrange sentence and write Story	Pearson Correlation	.440**	.220**	.347**	.181**	.381**	.543**	1	.268**	0.124	0.114	-.161*	-.177*	.393**
	Sig. (2-tailed)	0.000	0.002	0.000	0.010	0.000	0.000		0.000	0.078	0.108	0.022	0.012	0.000
	N	201	201	201	201	201	201	201	201	201	201	201	201	201

Q08 Write Paragraph	Pearson Correlation	.305**	.622**	.234**	.349**	.261**	.351**	.268**	1	.477**	.349**	.246**	.406**	.717**
	Sig. (2-tailed)	0.000	0.000	0.001	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000
	N	201	201	201	201	201	201	201	201	201	201	201	201	201
Q09 Write Story	Pearson Correlation	.259**	.553**	.218**	.268**	.183**	.274**	0.124	.477**	1	.529**	.475**	.405**	.755**
	Sig. (2-tailed)	0.000	0.000	0.002	0.000	0.009	0.000	0.078	0.000		0.000	0.000	0.000	0.000
	N	201	201	201	201	201	201	201	201	201	201	201	201	201
Q10 Describe the graph	Pearson Correlation	.224**	.337**	.258**	.279**	.237**	0.087	0.114	.349**	.529**	1	.427**	.158*	.641**
	Sig. (2-tailed)	0.001	0.000	0.000	0.000	0.001	0.217	0.108	0.000	0.000		0.000	0.025	0.000
	N	201	201	201	201	201	201	201	201	201	201	201	201	201
Q11 Write a letter	Pearson Correlation	0.010	.281**	-0.068	.202**	0.088	0.043	-.161*	.246**	.475**	.427**	1	.506**	.550**
	Sig. (2-tailed)	0.886	0.000	0.337	0.004	0.215	0.545	0.022	0.000	0.000	0.000		0.000	0.000
	N	201	201	201	201	201	201	201	201	201	201	201	201	201
Q12 Write a dialogue	Pearson Correlation	0.019	.423**	-.195**	0.107	0.061	0.113	-.177*	.406**	.405**	.158*	.506**	1	.500**
	Sig. (2-tailed)	0.786	0.000	0.006	0.130	0.393	0.111	0.012	0.000	0.000	0.025	0.000		0.000
	N	201	201	201	201	201	201	201	201	201	201	201	201	201
TOTAL	Pearson Correlation	.540**	.756**	.417**	.532**	.441**	.568**	.393**	.717**	.755**	.641**	.550**	.500**	1
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
	N	201	201	201	201	201	201	201	201	201	201	201	201	201
**. Correlation is significant at the 0.01 level (2-tailed).														
*. Correlation is significant at the 0.05 level (2-tailed).														

Table-12 (b): Statistics of total test items for measuring correlation

Interpretation: The following interpretation has been prepared based on the table no. 13

Q1(MCQ).

First, its correlation with q2 .341** , q3 .541** q4 543**, q6 592** q7 values with 440** and q8 305** are at the level of **(0.3-0.5).** Therefore, it has met the criterion 1

Second, correlation with q9 (259**), q10 (0.224**), q11 (0.010) and (0.019) for MCQ have failed to criterion 1 **(0.3-0.5).**

Third, the correlation with total is 0.540 which does not meet criterion 3 (**(above 0.7).**

Comments: The MCQ is question item stands at the positive weak uphill level in terms of validity. Hence, it requires a massive reconstruction to suit for this group.

Q2. Short Question:

First, short question with value 0.341** correlates with q4 0.327**, q6 0.453**, q9 0.553**, q10 0.337** and q12 0.423** match the criterion 1 (0.3-0.5) whereas its correlation with q8 0.622** stands at the moderate uphill level.

Second, short question with value 0.341** correlates q3 0.209**, q5 0.286**, q7 0.220** and q11 0.281** do not meet criterion 1.

Third, short question with value 0.341** can correlate with total score 0.756** which can meet criterion 3.

Comments: This item correlates 5 items at weak uphill positive level while the same does not function to correlate with 4 items as the values remain less than 0.3. On the contrary, its correlation a single item at the moderate uphill level and its correlation with the total meet demand of criterion 3. Therefore, this item requires a little revision to build its construct validity.

Q3 Read the Text and Fill the gaps:

First, its value 0.541** can correlate with the q4 0.396**, q6 0.467** and q7 0.347** which meet criterion 1 (0.3-0.5.

Secondly, its correlation with q5 0.178*, q8 0.234**, q 0.218** and q10 0.258** finds no linear relationship. Besides, its correlation q11 -0.068 and q12 -.195** indicates that functional power of this in group stands at weak downhill negative. Hence, values of the statistical calculation projects that this items do not meet the criterion 2.

Thirdly, its correlation with total values 0.417** miserably fails to meet the criterion 3.

Comments: This item does not meet the two major criteria. Hence it requires either deleted from the group or deserves massive revision.

Q4 Complete the table from the information of the passage:

First, q4 with its value 0.543** correlate q6 0.433** and q8 0.349** indicate that their relation stands at the weak uphill relationship. It shows that relation meets criterion 1 (0.3-0.5.

Secondly, its correlation with q5 0.206**, q7 0.181**, q9 0.268**, q10 0.279** q11 0.202** and q12 0.107 find no linear relation which indicates that the item does not meet the criterion 2.

Thirdly, its correlation with values 0.532** does not meet criterion 3 as well.

Comments: This fails two major criteria and its relation with other question is seriously low. Hence, the item should be deleted from this group.

Q5 Write summary:

Its value has the value of 0.268** which does not find any linear relation with rest of the question item. Besides, its correlation with q6 0.227**, q8 261**, q9 183** and q10 0.237** is statistically significant but finds no linear relationship and it does not meet criterion 1. On the contrary, its correlation with q11 0.088 and q12 0.068 are not statistically significant. Again its correlation with q7 0.381** is at the weak uphill level but it does not have the sufficient value to stand at least the weak uphill level.

Secondly, its functional value is so low that does not meet criterion 2 as well.

Thirdly, its correlation with total value 0.441** does not meet the criterion 3.

Comment: This item requires a massive redesign and reconstruction else it will not be able to bring out the expected writing skills of the students. If the redesign is possible, this items needs to be deleted from the question paper.

Q6 Match the sentences from the columns:

Firstly, as per criterion 1 the value of q6 stands at weak uphill level and its value also correlate with q2 0.453**, q3 0.467**, q4 0.433**, q7 0.543** and q8 0.352**which stand at the weak uphill level. They are statistically significant. Moreover, this item meets the criterion 1.

Secondly, its correlation with q10 0.087, q11 0.043 and q12 0.113 indicate that its relation with them is not statistically significant. Hence, the students find this item easy and it fails to distinguish between weak and meritorious students. Moreover, item does not meet the criterion 2.

Thirdly, its correlation with total values stand at weak uphill level which does not fill requirement of criterion 3 i.e. 0.7.

Comments: This item deserves structural redesign and requires a logical reconstruction. If it is not corrected properly, students will find it easy to answer and they will not develop their expected skills.

Q7 Re-arrange Sentences and make story:

First, its value 0.440** correlates with q3 0.347**, q5 0.381** and q6 0.543** and they are statistically significant. Its correlation meets criterion no.1. On the contrary, its correlation with q2 0.220**, q4 0.181** and q8 0.268* is statistically significant but its value does not find any linear relationship.

Secondly, its correlation with q11 -0.161* and q12 -0.177* bears a weak down hill (negative) trend which signifies that the item is highly predictable and extremely easy for students. It fails to meets the criterion 2.

Thirdly, its correlation with the total value 0.393** does not match with the criterion 3 as well. Hence, it can be assumed that the item is extremely easy.

Comments: This item does not deserve to be reformed for further use. Therefore, it requires immediate deletion form the group. It is also suggested that this item needs a massive reconstruction so that the expected skills to be brought out while students attempt it.

Q8 Write Paragraph:

Firstly, value of q8 0.305** correlates with 5 items out of 12 (q4 0.349**, q6 0.351**, q9 0.477**, q10 0.349** and q12 0.406**) which meet the criterion no. 1. its correlation is 99% statistically significant.

Secondly, its values correlates with a single item-q2 0.622** which meets criterion no. 2. On the other hand, its value finds no linear relation with q3 0.234**, q5 0.261**, q7 0.268** and q11 0.246**. Yet they are statistically significant.

Thirdly, its correlation with total value 0.717** meets criterion no.3.

Comments:

From the structural point of view the question item is well designed and it can distinguish between weak and strong students in terms of their merits. Hence, this function well in the group and for this we can restore this question item for question bank.

Q9 Write Story:

First, its value correlates 5 question items out of 12 (q2 0.553**, q8 0.477**, q10 0.529**, q11 0.475** and q12 0.405**), which meets the criterion no. 1 and they are statistically significant. On the contrary, its value does not find any linear relationship with 5 items and they are q3 0.218, q4 0.268**, q5 0.183**, q6 0.274** and q7 0.124.

Secondly, it is significant none of the question stands out to meets the criterion no. 2.

Thirdly, the value of q9 0.259** correlates with the total value of 0.755** which indicates that it meets criterion no 3 in a partial manner.

Comments: The question is moderately good in terms structural design but requires a little correction to improve its quality to stay in the group.

Q10 Describe the graph:

First, only two question items-q2 0.337** and q8 0.349** meet the criterion 1, while rest of question items-q3 0.258**, q4 0.279**, q5 0.237* find no linear relationship with q10 and they are all 99% statistically significant. Besides, q12 0.158* with its value is 95% statistically significant. Moreover, q6 0.87 and q70.114 find no linear relation with q10 and they are not even statistically significant.

Secondly, q9- a single question item meets criterion no. 2.

Thirdly, the total value of 0.641** does not meet criterion no. 3.

Comments: This question item does not have functional and logical strength to distinguish between weak and strong students in terms of merit. Hence, items claims massive revision else it can be deleted from the list.

Q11 Write a letter:

Firstly, q11 0.011 with its lowest out of all correlates only two question items-q9 0.475** and q10 0.427** and they are 99% statistically significant. They can meet the criterion no. 1. Despite the rest of three items- q2 0.281**, q4 0.202** and q8 0.246** statistically significant they harbor no linear relation with q11. On the contrary, the two items- q5 0.088 and q6 0.043 neither have linear relation nor they are statistically significant. Besides, q11-its correlation with q3 -0.068 and q7 -.161* is a weak downhill negative.

Secondly, none of items are found to stand out to meet the criterion no. 2.

Thirdly, its value correlates with the total 0.550** which does not meet the criterion no. 3.

Comments: Its correlation to meet the criterion no. 1 is partially fulfilled; hence it has a faulty structural design. Besides, its logical connection within the sentence structure is miserably low and its functional capacity in group extremely low as well. Therefore, this item needs to be deleted.

Q12 Write a dialogue:

First, q12-0.019 correlates with q2 0.423**, q8 0.406**, q9 0.405** and q11 0.506** which can meet criterion no. 1. On the other hand, q4 0.107, q5 0.061, q6 0.113 and q10 0.158* find no linear correlation with the said item. Besides, q10 is 95% statistically significant.

Moreover, q3 -.195** is 99% while q7 -177* is 95% statistically significant whereas both them stand at the level of weak downhill negative correlation.

Secondly, there is none to main a correlation which meet criterion no.2 which reveals a major weakness of this item in terms of logical connection within this item's ideas and also distinguishing power that can demarcate between weak and strong students.

Thirdly, its value correlates with total 0.500** which is 99% statistically significant but this correlation does meet criterion no. 3.

Comments: This is extremely a weakly built item which needs to be deleted.

Summary of the Validity Measurement

1. In case of Item to Item correlation (Single Item Vs Single), what is significant is that 5 out of 12 (q5, q9, q10, q11 and q12) question items are found to maintain No Linear Correlation whereas 4 out 12 question items stand at A Weak Uphill Positive Correlation. Besides, in the same case, only 3 out of 12 question items remain at the Moderate Level Correlation. It is urgent to address issue of validity that with regard to Item to Item Correlation that a greater number of Items hovers around at the No Linear Correlation and A Weak Uphill Positive Correlation level than that Moderate Level Correlation, A Strong Uphill Positive Correlation and A Perfect Uphill Positive Correlation.

Hence, we can assumed that there is a dearth of Construct and Content Validity which indicate that almost 5 out of 12 question items are not worthy of retention i.e. they are so

weakly designed that they should not belong to this group. Moreover, 4 out of 12 items require massive revision and reconstruction whereas the rest can be retained after a little modification to reach at strong uphill positive correlation and a perfect uphill positive correlation.

2. In case of Single item Vs Total Correlation, 6 (q1, q4, q6, q10, q11 and q12) out of 12 question items find their positions at the moderate level. On the other hand, in the same case, 3 (q3, q5 and q7) out of 12 items remain at the weak uphill positive correlation. However, the rest 3 items have been taken their positions at the level of A Strong Uphill Positive Correlation. For the case of Single item Vs Total Correlation, the expected correlation has been assumed that a majority number of items' Correlation would remain at least at the moderate level and the rest of the items' would hang around at A Strong Uphill Positive Correlation and A Perfect Uphill Positive Correlation. In a valid set of question neither of the items should pop up as the weak uphill positive level or less than moderate uphill positive correlation level in terms of Single Vs Total Correlation.

Summary of the Reliability Measurement:

1. In case of 12 items, the Cronbach's Alpha values reached $\alpha = 0.798$ and with respect to the measurement if deleted, q3 .794, q5 .792, q11 .795 and q12 .795 are very close to required reliability. Hence, they could be acceptable in terms Cronbach's Alpha reliability range. Similarly, the items-q1 .787, q4 .788 and q6 .783 remain a little away from acceptable of reliability but hover around the required range of reliability but question items- q2 .758, q8 .765, q9 .759 and q10 .779 do not even linger around required range of reliability. Therefore, there found a reasonable gap between existing values of the aforementioned items and the required Alpha range. Hence, there is dearth of reliability found in the said items.

2. On other hand, the Cronbach's Alpha α of q7 .801is (if deleted) is higher than 0.798, the item can be regarded as unreasonable item and it can be deleted from the group.

The final part of the article includes recommendations. The recommendations included here call for the attention of National Academy for Education and Management (NAEM) and Ministry of Education for further improvement of the test quality. Moreover, the suggestions and recommendation are meant for the setters and moderators to improve the professional quality and skills so that they can tend to more responsible in designing the test. The limitations of the study will help aspirant researcher to find their paths in the field of testing and evaluation.

Recommendations

1. Make sure that the question item should have appropriate level of difficulty level and discrimination Index so that the students cannot find the items too difficult and too easy.
2. All alternatives (Distractors) of MCQ should be plausible and arresting so that students can take time and strive to choose the right one. Importantly every distractor should have minimum tempting quality to be chosen by at least 5% students or in some cases chosen by 10% students.

3. Properly trained setters and moderators are assigned to setting and moderating and therefore, they can have ability to insert accurate level of validity and reliability so that the flaws related to item construction can be reduced a lot.
4. A short-term training on English Language Assessment should be designed for the test setters and moderators where all the issues regarding validity and reliability will be reflected while designing a test.
5. A committee of expertise on testing and evaluation needs to be formed to formulate the guidelines and principles of designing test items, test instructions and marks distributions. Moreover, they should develop separate question items based on the skills.
6. A national rubric should be provided to the examiners to minimize the range of marking discrimination from teacher to teacher.
7. A reasonable time should be given to the examiners and during this period work load should be decreased from these selected examiners to have the best output from them.

Acknowledgement

We sincerely thank the National Academy for Educational Management (NAEM) for funding this research. We also extend our gratitude to all the participants for their voluntary involvement, which was essential to the completion of this study.

References

Anderson, S. B. (1975). Encyclopedia of Educational Evaluation: Concepts and Techniques for Evaluating Education and Training Programs.

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

Ebel, R. (1972). Essential Educational Measurements. Prentice Hall, Inc. Englewood, Cliff. New York.

Garret, H. E. (1964). General Psychology. Eurasia Publishing House, New Delhi.

Hamp-Lyons, L (2000) Social, professional and individual responsibility in language testing, System 28, 579–591.

Hughes, A. (2003). *Testing for Language Teachers (2nd Edition)*. Cambridge: Cambridge University Press.

Hyland, K. (2006). *English for Academic Purpose: An advanced resource book. Routledge.*

Khan, R. (2010). English language assessment in Bangladesh: Developments and challenges. *Language assessment in Asia: Local, regional or global, 121-157.*