Global Language Review (GLR)
URL: http://dx.doi.org/10.31703/glr.2016(I-I).05

Test-Score Reliability of Essay Questions in BA Examination

Muhammad Asif Khan*

Umbreen Ishfaq†

* Assistant Professor, Department of English, The Islamia University of Bahawalpur, Bahawalpur, Punjab, Pakistan. Email: asifkhan04@yahoo.com

† Assistant Professor, Department of Education, The University of Haripur, Haripur, KP, Pakistan.

- **p-ISSN:** 2663-3299
- **e-ISSN:** 2663-3841
- **L-ISSN:** 2663-3299
- **Vol:** I (2016)
- **Page:** 58 – 65
- **DOI:** 10.31703/glr.2016(I-I).05

Abstract *Writing is one of the basic language skills. Both objective tests and essay questions can be used to evaluate students' performance in writing skills. objective tests are generally considered useful as they help in measuring certain abilities accurately and they yield reliable scores. But essay tests can be used more effectively to measure certain complex learning outcomes such as organization, synthesis, and analysis. Essays provide an effective form of testing to evaluate the communicative skill of students at the higher level. Such items have the problems of consistency, objectivity, and reliability of their scores. The present study focuses on the issue of scorer reliability and attempts to suggest an objective model of scoring for the free-response essay question set in the B.A. examination for Islamia University of Bahawalpur in the subject of English compulsory.*

Key Words:

Language Testing, Writing Skill, Scorer Reliability, Essay Question

Introduction

Writing compositions, especially essays is an important part of the English language learning process in Pakistan. Both personal experience and observation show that the effectiveness of the composition and essay type questions is not properly utilized. Almost everything from teaching how to write essays to the construction of the essay-type question to its marking and grading involves problems. Essay questions are used to measure such abilities and skills as the objective test items cannot measure effectively. Essay questions help in achieving important learning objectives such as the ability to recall, select, organize, and express ideas, and the ability to provide rather than identify information. So, the essay questions are used for measuring complex learning objectives. Essay questions give students a degree of freedom. Students get the freedom of response,

of selecting, relating information, and presenting ideas in their own words. This freedom, though it makes essay questions valuable, also creates difficulties. For example, the essay item construction requires great care, it must contain validity, then, evaluating the answers is a difficult and time-consuming task, etc. However, the learning objectives achieved through essay questions are so valuable that the difficulties that such questions involve can be tackled. The present study focuses on the difficulty of obtaining scorer reliability. The unreliability of the scoring is perhaps the most serious limitation of the essay questions. Different teachers give different scores to the same answers to essay questions, the scores of even the same teachers vary if the scoring in done at different times. So there are great variations in scoring the same answer. The present study is an attempt to suggest an objective and reliable model of scoring for the fifteen marks essay question which the students of The Islamia University of Bahawalpur are required to attempt in the B.A. examination. In the subject of English (Language) which the students are required to pass as a compulsory subject, the students are asked to choose one item from a list of usually five extended-response essays. The present is an attempt to identify the elements/ components of the essay and to objectify the scoring of the free-response essay questions by suggesting a pattern to achieve inter-scorer reliability.

What is an essay?

Answering the question, 'what is an essay item?' Tom Kubiszyn and Gary Borich (1990:98) write that an essay "...demands that the student compose a response, often extensive, to a question for which no single response or pattern of responses can be cited as correct to the exclusion of all other answers."

Essays: Their learning outcomes:

Essay questions may be more useful than objective items as they help in measuring certain complex learning objectives effectively. Kubiszyn and Borich (1990:101) list some of the learning outcomes for which essay questions may be used:

- Analyze relationships
- Arrange items in sequence
- Compare positions
- State necessary assumptions
- Identify appropriate conclusions
- Explain cause-and-effect relations
- Formulate hypotheses
- Organize data to support a viewpoint
- Point out strengths and weaknesses
- Produce a solution for a problem.

- Integrate data from several sources
- Evaluate the quality or worth of an item, product, or action
- Create an original solution, arrangement, or procedure

Many of the abilities which essay tests, it is claimed, help to measure, such as analysis, creativity, synthesis, organization and judgment are quite complex in nature and are not clearly defined. Essay tests also, just like any other from of testing, have their advantages and disadvantages. Gilbert Sax (1997, 119-122) gives the following strengths and weaknesses of the essay tests:

Advantages of Essay Tests:

i. They allow free response:

This is perhaps the greatest advantage of this type of tests from the students' point view.

ii. They eliminate guessing:

Because the students are required to recall and supply rather than select from the given information.

iii. They are practical:

Compared with multiple choice tests, the essay tests are easier to prepare and score, and take less time. That is why many teachers prefer them.

iv. They reduce assembling time:

They are easier to administer and conduct.

v. They can measure divergent thinking:

Because they give great freedom of response, they provide more chances of unusual responses reflecting divergent thinking.

Disadvantages of Essay Tests:

Essay tests have some serious flaws:

i) They are difficult to score:
While they are easier to score, they are difficult to score objectively.

ii) They measure limited knowledge:
A multiple choice test with greater number of items can measure wider area of knowledge than does an extended essay which measures limited aspects of knowledge.

iii) They are time-consuming:
They are time-consuming since students spend much time answering only one extended-essay question and the teachers spend many hours in reading lengthy essays.

iv) They are subject to bluffing:
Students with poor preparation try to get passing marks by writing something, even if it is irrelevant.

v) They typically require rote memorization:

An essay topic anticipates creativity and originality. Practically, few essay questions require creative thinking, and most essay topics typically require the lengthy presentation of memorized material.

Despite their limitations, essay questions continue to provide a useful form of testing which, especially for the non-native learners of the English language at the higher level, effectively measures certain complex learning outcomes by offering an extensive exercise in the writing skill. The writing of essay items requires great care and skill. The poorly constructed essay item can be quite problematic. It may encourage the student to use memorized material. It may even make it difficult for the student to know what response is required. It may, in turn, make the accurate and objective scoring difficult for the rater.

Extended and Restricted Response Essays:

Here, it may be useful to make a distinction between two prominent types of essay question. An extended-response essay item gives the student the freedom to decide upon the length, scope and complexity of the response. This type of essay item is useful for measuring the cognitive abilities such as synthesis and evaluation. Ideally, an extended-response essay item should be given as a term paper assignment or a take-home test. A restricted-response essay item assigns a specific problem/ task and the student is required to express himself / herself and give information within the limits of the proposed problem. The boundary between the extended and restricted response essay items is a minute and vague one. David Allan Payne (2003:232) writes, "No hard and fast rules determine when a 'restricted response' essay becomes 'extended'."

What is Reliability?

Reliability is one of the characteristics required for a good test. Peter W. Airasian (2001, 253) writes: "Reliability is concerned with the stability and consistency of assessments - e.g., are the results typical of a particular pupil's performance?" W. James Popham (2006, 115) observes that reliability is synonymous with consistency. He writes: "If a test is unreliable (that is, if it measures whatever it's measuring with inconsistency) there is little likelihood that educators can make valid inferences about the meaning of students' test performances." An unreliable test can be a waste of time and effort for both teachers and students since its conclusions cannot be trusted.

Test and Scorer Reliability:

There are several forms of reliability, but two deserve special attention here. Test reliability is affected by many factors which include; the adequacy of the samples

of students' performance, the conditions under which the test is taken, student motivation, weaknesses in the test, etc. "Scorer reliability is the extent to which different observers or raters agree with one another as they mark the same set of papers" (Sax, 1997, 285). Scorer reliability refers to the consistency with which tests are evaluated. Scorer reliability is high if one scorer gives the same score repeatedly for the same response, and if two or more scorers give equal scores for the same performance.

Scorer reliability is not an issue in the case of objective tests, but, it becomes a matter of great importance in the case of compositions and essay tests, especially the free response essay items. The low reliability of the essay test scores is the most serious limitation of this type of tests. According to Robert L. Ebel and David A. Frisbie (1991, 191), overall, three conditions are responsible for this low reliability: (i) the limited sampling of the tasks, (ii) the vagueness of the tasks, and (iii) the subjectivity of the scoring. In conditions when the scoring of tests involves subjectivity, interjudge (interscorer) reliability and / or intrajudge reliability become important. Gay (1991, 172) writes: "Interjudge reliability refers to the reliability of two (or more) independent scorers; intrajudge reliability refers to the reliability of the scoring of individual scores."

Methods of Scoring Essays:

The scoring of essay tests can be made objective by careful planning in scoring. There are two prominent methods of scoring the essay tests:

i. Holistic Scoring or The Impression Method: Sax writes: "In holistic scoring,...the teacher, rather than examining every sentence or main idea to determine how many points the student is to receive, estimates the overall quality of each paper" (1997, 132). It is the scorer's overall impression of the response as a whole that is important. The holistic approach is useful when a lot of papers are scored by multiple raters.

ii. Analytic Scoring: This is an especially useful method of scoring for a large number of limited-response essay questions. The scorer decides how much weight each task or learning outcome will have and informs students about it. The scorer using this method evaluates specific categories. Gay (1991, 226) observes: "The analytical approach involves identifying all of the aspects or components of a perfect answer and assigning a point value to each." This method of scoring is objective and reliable.

The elements/ components of writing the essay:

It is not easy to determine the elements of the writing skill. Generally, an essay item measures two major skills: the ability to use language, the communicative

skill, and the ability to produce relevant and convincing ideas and arguments. The essay question given to the B.A students, basically, evaluates their communicative skill, the content, for most of the essay topics, usually requires the students to reproduce memorized material. It is a bit unfair to expect from students originality and creativity, qualities which are quite rare and which can be found only in persons of the highest intellectual caliber. The discussion of the elements or components of the essay questions involves two key aspects: determining the components, and deciding what weighting each will get, i.e., deciding whether all components will get equal importance or some will get more percentage than the others. The rating scale offered by Paul Diederich, for example, weighs organization 50 percent, style 30 percent, and mechanics 20 percent (cited in Payne, 2003:245). For practical purposes, the elements of the essay item are discussed under five heads:

i. Grammar: This refers to the correct use of different parts of speech: nouns, pronouns, adjectives, verbs, prepositions, articles, etc.

ii. Structures: This refers to the variety of sentence structure, the ability to use different kinds of sentence according to function and clause structure (or sentence length),

iii. Vocabulary: It is concerned with the choice and range of lexical items,

iv. Mechanics: It refers to the use of spelling and punctuation, and,

v. Content and Organization: That is, ideas and the development of the discussion.

Scoring the Essay Question: A Proposed Model:

In order to make the scoring of essays objective, it is suggested that all elements (mentioned above) should be given equal importance. Each component can be evaluated under three broad categories: poor, fair, and good. Each component may get a maximum of three points, one for poor, two for fair, and three for good performance. The following table shows the marking scheme:

Table 1. A Proposed model of scoring the essay.

	Poor (1 Mark)	**Fair (2 Marks)**	**Good (3 Marks)**
Grammar			
Structures			
Vocabulary			
Mechanics			
Contents			

A hypothetical performance that is poor in content, fair in structures and vocabulary, and good in grammar and mechanics will be scored like this:

Table 2. The scoring of a hypothetical performance.

	Poor (1 Marks)	**Fair (2 Marks)**	**Good (2 Marks)**	**Marks Obtained**
Grammar				3
Structures				2
Vocabulary				2
Mechanics				3
Contents				1
Total	**(Out of 15)**			**11**

Suggestions for scoring the Essay Tests:

i) The essay item should be constructed carefully. Topics which are quite vague or too broad in scope should be avoided. The examiner should write unambiguous essay items. Students should not be given essay topics which they do not understand clearly. If the student can interpret the question in a way which the examiner has not anticipated, it means that the test is not a reliable one.

ii) The examiner should make the writing tasks clear to the student. The student should not be allowed too much freedom to write as he/ she chooses to write. A good essay item makes clear to the student what he/ she should write and what he/ she should avoid.

iii) The examiner should provide clear instructions to the students.

iv) The essay topics should be interesting and motivating for the students. They should not require the student to write on a topic that is beyond the knowledge, understanding, and experience of the student.

v) The essay topics must not be reused. The topics that are repeated in the examination encourage the students to reproduce memorized material.

vi) If a large number of student performances have to be scored by multiple examiners, it is important that the scorers should first agree on which aspects of the essay will be evaluated and what weighting each aspect will get.

vii) The students should be made aware of the format of scoring. The classroom tests can be quite helpful in this regard.

If the suggestions given above are followed, the scoring of the essay question can be made objective.

References

Airasian, P.W. (2001). *Classroom Assessment: Concepts and Applications*. 4th Ed. New York: McGraw-Hill.

Ebel, R.L., & David A. Frisbie. (1991). *Essentials of Educational Measurement*. 5th Ed. New Delhi: Prentice-Hall.

Gay, L.R. (1991). Educational Evaluation and Measurement: Competencies for Analysis and Application. 2nd Ed. New York: Macmillan.

Kubiszyn, T., and Gary Borich. (1990). *Educational Testing and Measurement: Classroom Application and Practice*. 3rd Ed. USA: HarperCollins.

Payne, D, A. (2003). *Applied Educational Assessment*. 2nd Ed. Belmont, CA: Wadsworth/Thomson Learning.

Popham, W.J. (2006). *Assessment for Educational Leaders.* Boston: Pearson.

Sax, G. (1997). *Principles of Educational and Psychological Measurement and Evaluation*. 4th Ed. Belmont, CA: Wads-worth.