# Opinion Pieces

## THE CASE AGAINST ESSAY TESTS

Robert C. Sorensen[1]
University of Nebraska
Lincoln, NE 68583-0914

### Abstract

Essay tests are a part of many and varied types of educational programs. But these tests are subject to several educational and logistical problems. In this philosophical paper, these problems are identified and discussed. Appropriate use of essay tests is described. Suggestions are offered for addressing inherent problems in essay tests, or selecting a more appropriate form of performance assessment. Secondary references are provided which should give the reader with access to the voluminous literature of test design, quality, selection, and utilization.

### Introduction

Tests are ubiquitous and pervasive parts of the academic experience. Although their main function is to measure student accomplishment, they also serve as active learning experiences, sources of motivation, indicators of the effectiveness of instruction, and opportunities for students to practice coping skills for stressful situations. Their contributions to educational success are difficult to measure but are surely considerable. Therefore tests are very important in the educational enterprise and testing must be done well if students are to get the most out of their educational experiences.

Experience has suggested that few teachers have completed a course in tests and measurements. Thus the selection and design of a testing program is often based on individual experience, hearsay, and "folk wisdom", the latter two often including a certain amount of mythology. Despite this situation, these sources of information have served us remarkably well in many cases. In other cases, however, evaluation of learning is often ineffective, or inefficient, or both, and may even be seriously biased. Inappropriate selection of a performance evaluation program is often at fault. When necessary background and experience in performance evaluation are lacking, the essay test format is likely to be chosen by default.

A view common among teachers is that essay tests are superior to all other kinds of written tests, and that other subjective and all objective tests are used only as poor substitutes whose selection is made necessary by some characteristic of the testing situation. This view is patently false (Mehrens and Lehmann, 1978, Pg. 211). Although essay tests have their place, other types of assessments have many advantages and are frequently superior to essay tests. My objective in this paper is to discuss the serious shortcomings of essay tests and suggest some ways to improve them or to identify alternatives. It is not my intent to discourage the use of essay tests where they are well-suited to the task.

I do not intend to provide a comprehensive review of the voluminous literature relating to the use of tests in education. In order to avoid distracting the reader with copious literature citations, I have chosen to use secondary references. This practice should maintain the focus on the points in the paper, yet provide easy access to the literature on points of interest.

### Bloom's Taxonomy of Educational Objectives

Any discussion of tests must consider the cognitive levels of Bloom's Taxonomy (Bloom, 1956). Since most teachers are familiar with this taxonomy, I treat it only briefly. More detail can be found in Jacobs and Chase (1992), Pg. 17-19. Bloom lists six levels of cognitive activity, which I paraphrase. The first is knowledge, simple recall of information. Second is comprehension, recall of information in a slightly altered context. The third is application, the ability to apply information to new situations. Fourth is analysis, dividing a system into its component parts. Fifth is synthesis, creating

[1]Professor of Agronomy, Emeritus

new ideas based on previous concepts. Sixth is evaluation, deducing the value, effectiveness, or quality of a proposal. Bloom suggests that as one moves from knowledge to evaluation, increasingly demanding levels of cognition are required. A good practice for any test is to prepare a Table of Specifications for the course objectives (Jacobs and Chase, 1992, Pg. 20-24; Mehrens and Lehmann, 1978, Pg. 174-180) which specify how much of the test is devoted to each of the cognitive levels. This table then provides a guide for test construction.

## Background

As a basis for this discussion, I believe most teachers agree on several points. First, tests are not the only way, and in many cases, not the best way to measure academic performance. As we compare tests, we must also consider papers, oral presentations, design projects, case studies, written recommendations, and several other types of assessments. Although some of these methods fall prey to some of the same problems as essay tests, they may have much to offer in given situations. Second, all tests have strengths and weaknesses. In each situation, there are usually more than one type of test which can be used successfully. However, some are more effective or efficient, or both, than others.

Third, essay tests stand alone among tests in being able to measure cognition at the synthesis level. Other types of tests appropriately designed, can be effective at all other levels. Synthesis can be measured effectively, however, by several other assessment activities such as papers, oral presentations, and design projects, which are frequently more effective in measuring overall student accomplishment than essay tests.

The primary purpose for most course tests is to measure the degree to which students have attained the course learning objectives. Fairness, consistency, and effectiveness of instruction require that if a student is to be tested on some knowledge, skill, or ability, it must be identified specifically in detailed, measurable, course learning objectives (Mehrens and Lehmann, 1978, Pg. 181-183). Fourth, there must be a clear relationship between the test items and the objectives. Student requests for old tests are much more effectively satisfied by provision of good course objectives for a number of logistical and educational reasons than by distributing previous tests.

Fifth, comparison of test items should always be based on well-prepared examples. Nothing is served by comparing well-written questions of one type with poorly-prepared or inappropriate questions of another type. It is consistently true that good test items of any type are not easy to prepare. Still, we must be aware of dependable indicators of question quality.

As a basis for the following discussion, I must define what I mean by a quality essay test question. Since student accomplishments at other cognitive levels can usually be addressed more effectively and efficiently by other test types, the good essay question should require synthesis of material to produce something new. The answer may be short or long, simple or complex. If it is a good essay question, it will probably educe comments from memorization-oriented students such as "How are we supposed to know that?" or "When was this covered?". The question must address a single issue and be very clearly written. It is likely to have a variety of worthy answers depending on how the student approaches it. Following are some short essay questions which satisfy the above requirements:

1. Describe the economic structure of the family farm of the future (2010-2020).

2. It has been said that genetically modified organisms in food crops (frankenfood) will lead to problems in the future. Is this statement true or false? Why do you think so?

3. Give five reasons why a diet including meat is superior to a vegetarian diet for teenagers.

The wording of these questions is subject to improvement. We must assume that specific answers to these questions were not provided to the students in the course lectures or materials.

## Inherent Problems of Essay Tests

The most serious problem of essay tests is their low reliability in scoring. The credit given on a question may vary greatly from one evaluator to another, one time to another, and one test form to another. In workshops I have done, scores assigned on the same essay test varied from 40% to 90% among the teacher participants. This problem is addressed in almost every tests and measurements text, and I refer the reader to them for documentation and additional information. Good sources are Ebel and Frisbie (1986), Pg. 129, Erickson and Wentling (1976), Pg. 121, Jacobs and Chase (1992), Pg. 106, and Mehrens and Lehmann (1975), Pg. 208. Each of these texts points to many research studies in support of this point.

The second problem with essay tests is one of efficiency and thoroughness. The amount of time required for a

student to write down an answer may severely restrict the number of teaching points which can be evaluated in the test. The better the essay question and the more capable the student, the more serious this problem becomes. A testing program in which a relatively low percentage of the teaching points in the course are evaluated by tests or other assessment procedures is not fair to the student and gives a biased view of the effectiveness of the course. This problem also tends to reduce test reliability. Jacobs and Chase (1992), Pg. 109, discuss this problem of limited content sampling.

Third, I have found no objective, reliable, and simple means to assess the quality of an essay test. Whereas objective tests can be evaluated by a variety of forms of item analysis, for essay questions personal and peer subjective evaluations are about all that is available. Similarly, there is no simple way a teacher can get a summary of student answers on an essay test. Tests and measurements texts usually do not address this problem.

Fourth, the preparation of a student's answer on an essay test may be adversely affected by poor writing ability, inadequate background, learning style, personality type, stress, and other factors. Although all tests are subject to influences by some of these factors, the essay test is most seriously affected. Some teachers intentionally include one or more of these factors in the scoring. For example, some teachers consider writing ability in the scoring of students' answers. This is certainly permissible if it is clearly stated in the course learning objectives and especially if remediation is provided.

Fifth, because of the freedom given to students to develop their answer in their own way by the essay format, it is very difficult for the teacher to give good feedback when pressed for the "right" answer. Many students are frustrated when the content of the full-credit answer cannot be clearly specified. Yet one of the strengths of the essay format is that a variety of "right" answers are likely. This is obviously a greater problem for beginning or less sophisticated students, and may adversely affect their motivation in the course.

Sixth, the scoring of good essay questions is very time-consuming. For well-designed questions used in appropriate settings the increased effort of performance assessment may be rewarded. But where there is any question about the efficacy of the essay format, machine scoring of objective tests is essentially error-free, and considerably more resource-efficient. An item-analysis of the questions should also be available.

The last problem may provide the most room for debate. I propose that the essay format is not consistent with most life experiences. In life, we are rarely asked to write out a discussion of what we think about a situation, particularly under time constraints and without access to external resources. We are more likely to be called upon to make decisions. Decision making is usually more effectively tested by good multiple choice questions in various formats than by essay questions where no decisions are required. If problem-solving skills are being measured, some type of guided design or case study exercise is usually more effective than written tests.

## Other Considerations in Testing

The most extensive problem with essay questions is their use where other types of questions or other assessment methods are clearly more effective and efficient. For example, essay questions should not be used at the knowledge or comprehension level (Jacobs and Chase, 1992, pg. 111, and Mehrens and Lehmann, 1978, pg. 213.) However, in my 35 years of teaching experience, a large part of the essay questions I have seen would be judged at this level.

As indicated previously, the cognitive level of a question is sometimes related to what was done previously in the course. For example, given the following question from plant science:

> Select a tillage system that would sequester the greatest amount of carbon from the atmosphere and justify your answer.

At first glance, this appears to be a question posed at the highest cognitive level, evaluation. But if tillage systems were compared in the lecture, handouts, or reading, and the most effective one identified in one of those sources, the question is simply a matter of knowledge, the lowest level. Even if the most effective tillage method were not identified in the course, the ability of the student to evaluate the situation could be learned much more efficiently from a multiple choice question with carefully designed distractors. For the reasons previously given, other types of questions are usually as good or better than essay questions at the levels of application, analysis, and evaluation.

The selection of a test program should be based on a thorough examination of the course objectives, student characteristics, the classroom environment, and abilities of the test writers. Student characteristics include background knowledge and experience, educational sophistication, attitudes and interest, and career aspirations. The classroom environment includes time available, equipment and

resources available, seating arrangements, and class format. The abilities of the test writer include knowledge, experience, and resources available. Assistance with selection of a written test format is available in many test and measurements texts. One of the best, adapted from Thorndike and Hagen (1969), is given in Mehrens and Lehmann (1978), Pg. 188.

Time constraints are a problem with essay tests, and may cause serious difficulties for some students. Therefore if essay tests are to be used, some means to alleviate this problem should be sought. Take-home tests, or an alternative period without time limits are examples.

There are many practices that can be used to improve to some extent the reliability of grading. These include multiple graders, grading one question on all tests before going to the next question, and specifying beforehand what the components of the answer should be. The latter, of course, tends to invalidate a major strength of synthesis, that is to allow creativity in the answers. Ebel and Frisbie (1986), Pg. 134 suggest ways to improve the reliability of essay tests.

Many teachers and some authorities (Ebel and Frisbie (1986), Pg. 126; Erickson and Wentling (1976), Pg. 121) say that essay questions are easy to prepare. This may be true of essay questions inappropriately used to test lower levels of cognition. It is usually not true for essay questions used appropriately for synthesis. Students must be given clear and specific guidelines in the question to direct them to what is expected in the answer. There is little hope of making any kind of unbiased, reliable assessment of what has been learned from a question that is too broad and ambiguous. Effort needed for preparing a good essay question is essentially equal to the effort needed to prepare good multiple choice questions. If this is not true, essay questions are probably constructed at low levels of cognition. However, since fewer questions are prepared, this is a net gain of time in favor of essay questions.

I believe that teachers place too much reliance on tests, in general. There are often other assessment methods which are equally or more effective. This is particularly true for essay tests which are used for assessing synthesis. Various kinds of writing, speaking, and design exercises are often more effective than essay tests. If volume of scoring is a problem and students are advanced, peer evaluation of student work can be a good learning experience when using these exercises (Sorensen, R.C. and M.S. Wilhite. 2000. Experiences with peer evaluations of student papers. NACTA J. [In press]). Some change in course design may be necessary to fit them into the course.

## Myths About Essay Tests

One myth, that good essay tests are easy to prepare, has already been addressed. A second is that essay tests improve student writing. This may be expected if writing improvement is a course objective and if specific and detailed assistance is given to students to improve their writing. One does not improve one's golf game or bowling score by practicing the same old mistakes. Unless there is purposeful intervention, little improvement is likely to occur. The same is true of writing. Jacobs and Chase (1992), Pg. 109, discuss this issue under the heading, "Essays often promote poor writing skills." Ebel and Frisbie (1986), Pg. 128. also point out possible negative effects of essay tests on student writing.

Ebel and Frisbie (1986), Pg. 127, propose that essay tests allow teachers to deduce student thinking patterns. However experience has taught me that there is simply no way that a teacher can know what was in students' minds by reading their answers. Many psychological and environmental factors affect what gets written on a test paper. Serious errors can be made by trying to extrapolate students' answers to describe what they were thinking about. This assumption also contributes to reduced scoring reliability. There is also some question whether deducing student thinking patterns is a worthwhile objective. In most avenues of life, performance is more important than process.

It has been suggested that essay tests are more likely to cause students to think than objective tests. I believe the opposite is true. My experience has been that under the usual press of time in an essay test, the student's attention is usually directed to getting as much information written down on paper as possible leaving limited time for thinking about, and organizing what is written. Students believe, and they are usually right, that this approach maximizes credit. This behavior demonstrates also why trying to deduce thinking patterns by reading student answers is so hazardous, as described above. Conversely, when the only mechanical requirement in an answer is writing or circling a letter, or blackening a circle, much more time is available for thinking about the question and justifying the answer.

Some will say that subjective tests in general and essay tests in particular are best because they give no clues to the answer. What is wrong with clues? Sometimes impenetrable memory blocks which would be fatal on an essay test can be breached by one simple clue in a distractor for a multiple choice question. In all our life we rely on clues to key our knowledge and fuel our inspirations. Why should tests be any different?

## Recommendations

Based on the ideas presented above, six recommendations for successful testing when considering essay tests arise:

1. Write quality course learning objectives since they are the basis for all assessment methods.
2. Do not use essay tests where other tests are more effective, or efficient, or both.
3. If essay questions are appropriate, take time to design them well and be prepared for a variety of acceptable answers.
4. If synthesis is required in attaining a course objective, consider methods other than tests.
5. If essay tests are used, incorporate methods to improve scoring reliability and be prepared to address their other shortcomings. If these methods cannot be identified, use a different type of test or assessment procedure.
6. Take or audit a university course in tests and measurements.

## References


Bloom, B. S. (Ed.) 1956. Taxonomy of Educational Objectives, Handbook I: Cognitive Domain. Longmans. New York, NY.

Ebel, Robert L. and Donald A. Frisbie. 1986. Essentials of Educational Measurement. Fourth Edition. Prentice Hall, Inc. Englewood Cliffs, NJ.

Erickson, Richard C. and Tim L. Wentling. 1976. Measuring Student Growth. Allyn and Bacon, Boston, MA.

Jacobs, Lucy C. and Clinton I. Chase. 1992. Developing and Using Tests Effectively. Jossey-Bass Publishers. San Francisco, CA.

Mehrens, William A. and Irvin J. Lehmann. 1978. Measurement and Evaluation in Education and Psychology. Second Edition. Holt, Rinehart, and Winston. New York, NY.

Thorndike, R.L. and E. Hagen. 1969. Measurement and Evaluation in Education. 3rd Ed. John Wiley and Sons, New York, NY.


***NACTA** will be 50 years old in 2004*

*If you have reminiscences or memorabilia of our first 50 years that you would like to share with the organization please notify **Dr. Wayne Banwart, NACTA Historian or Dr. Bob Gough, NACTA Editor.***