IDEA PAPER No. 17

center for FACULTY EVALUATION & DEVELOPMENT
Division of Continuing Education
Kansas State University

January 1987 KANSAS STATE UNIVERSITY

Improving Essay Tests

by
William E. Cashin
Kansas State University

"... it seems clear, even to the casual observer, that essay examinations still are widely used in spite of more than half a century of criticism by specialists in educational measurement ...
Coffman (1971, p. 271)

Like every assessment technique, essay tests have their advantages and limitations. The position taken in this paper is that essay tests, despite their limitations, have a number of strengths and, therefore, appropriate uses in higher education, as long as we are aware of the limitations. There is considerable agreement in the educational measurement literature about how essay tests can be improved. Many of these recommendations are based on experience; some of these recommendations are based upon empirical research. I have not cited the original research; interested readers can find the references in the standard educational measurement texts identified in the "References and Further Readings" at the end of this paper.

What Is an Essay Test?

Coffman (1971, p. 271) describes an essay test as "one or more essay questions administered to a group of students under standard conditions for the primary purpose of collecting evaluation data." Their scoring requires expert judgment rather than the application of a clerical key. Administration under standard conditions distinguishes the essay test from the term paper or project report. However, I suggest that many of the recommendations concerning essay tests can also be applied, with appropriate adaptations, to term papers, project reports, oral exams, and other student products or processes used in assessing student achievement, including mathematical problems and artistic productions.

Essay questions are often divided into two types: extended response questions and restricted response questions.

Extended Response Questions. Other than stating the topic, extended response questions leave students free to determine the content and to organize the format of their answer. The students decide which facts are pertinent, and how to organize, synthesize, and evaluate them. Perhaps the classic example is: "How I Spent My Summer Vacation." Such questions are most appropriate when our objective is to test writing (composition) skills, including conceptualization, organization, analysis, synthesis, and evaluation, giving the student maximum choice regarding topic.

Restricted Response Questions. These limit both the content and the form (e.g. describe vs. compare and contrast) that the student's answer may take. Most writers agree that restricted response questions are the appropriate form when we wish to test content. All of the examples of essay questions which follow will be examples of restricted response questions unless otherwise stated.

Strengths of Essay Tests

Essay tests have a legitimate place in higher education because of the following strengths:

1. ***Can test complex learning outcomes*** not measurable by other means. An obvious example is the ability to express oneself in writing.
2. ***Can test thought processes,*** the students' ability to select, organize, and evaluate facts, ideas, etc; and their ability to apply, integrate, think critically, and solve problems. (Note: all of these can also be tested by appropriately designed multiple-choice items—see IDEA Paper No. 16, *Improving Multiple-Choice Tests,* Clegg and Cashin, 1986.)
3. ***Require that students use own writing skills;*** the students must select the words, compose the sentences and paragraphs, organize the sequence of exposition, decide upon correct grammar and spelling, etc.
4. ***Pose a more realistic task*** than multiple-choice and other "objective" items. Most of life's questions and problems do not come in a multiple-choice format, and almost every occupation, including engineering, business, technical, and service jobs, requires people to communicate in sentences and paragraphs, if not in writing, at least orally.
5. ***Cannot be answered correctly by simply recognizing the correct answer;*** it is not possible to guess. (Students can bluff, however.)
6. ***Can be constructed relatively quickly.*** This advantage is short-lived because any time saved in constructing the test is lost when scoring it. All well constructed tests require time and effort; the only choice is in when these will be expended.

Limitations of Essay Tests

The focus in this paper is on using essays for assessment. When essays are used as a learning experience to provide the students an opportunity to exercise a skill and then to give them feedback about their achievement, the limitations described below are of less concern. However, **as an assessment technique,** essay tests have the following serious limitations.

5429-5191-887-10M

1. ***Only limited content can be sampled.*** Therefore, essay tests are unreliable in assessing content. Because answering essay questions takes more time than answering "objective" items, less content can be tested. Most exams only sample a very small portion of the domain of content and skills to be learned. Therefore, when we rely solely on essay tests, differences in students' scores will to some extent reflect the "luck of the draw"—the questions you happened to include on the test—as well as reflect differences in the students' command of the entire domain of what you were trying to teach.

2. ***Yield unreliable scores.*** Not only have studies found differences in the grades assigned to essay questions by different scorers, but they have found differences for the same scorer grading the same question at different times. Thus, differences in student grades on an essay test may be due to *who* scored the question, or *when* it was scored, in addition to *what* the student knew or wrote.

3. ***Scores can be influenced by the scorer's impression of the student,*** e.g., general impression of the student: halo effect and test-to-test or item-to-item carryover—knowing how well the student did on the previous test or item. Obviously, multiple-choice tests do not have this limitation.

4. ***Scores may be influenced by factors extraneous to the content being tested,*** e.g., handwriting, writing skills, spelling, and grammar.

5. ***Essay tests often provide the students with an opportunity to exercise POOR writing skills.*** When one considers the time pressure and anxiety connected with the typical essay test, it is surprising that the students do as well as they do. Most of the students' time in an essay test is spent physically writing. There is limited time to think, to organize creatively, to write a second draft, or proofread.

6. ***Essay tests are time consuming to score.*** Anyone who has ever graded essays needs no proof of this beyond his/her own experience.

Recommendations

These recommendations are divided into three sections: when should essay questions be used, constructing the test, and scoring the test.

When Should Essay Questions Be Used?

These recommendations are adapted from Ebel and Frisbie (1986).

1. ***To test writing skills.*** Obviously, the most appropriate way to test the students' ability to express themselves in writing is to have them write something (remembering that essay tests are less representative of day-to-day writing tasks than are papers, project reports, keeping a journal, etc.).

2. ***To test a small group.*** Despite all of the advantages of multiple-choice and other "objective" type items, when testing small groups of students, developing such items is not worth the effort. Short answer questions, e.g., one to a few sentences identifying or defining questions, can be useful to serve in place of multiple-choice items.

3. ***When the time to construct the test is more limited than the time to score it.*** Testing is a teaching responsibility so we have a professional obligation to plan ahead for it. However, constructing a make-up exam for one or a few students who were legitimately unable to take the regular exam would be an instance where the instructor would have limited time.

4. ***When the instructor has more confidence in his or her ability as a critical reader than as an "objective" test constructor.*** Granted that college teachers, like the rest of humanity, differ as individuals, nevertheless, I would suggest that college teachers should have in their repertoire the basic skills of their profession including the ability to construct reliable and valid tests, both "objective" and essay tests.

5. ***To encourage students to explore attitudes*** more than testing for cognitive achievement. This suggestion focuses more on teaching (helping the students learn) than on testing, but fits into our broader approach suggesting that readers consider these recommendations not just for essay tests in the narrow sense, but also for papers, reports, journals, etc. Furthermore, in a later IDEA Paper it will be urged that testing, and all of our assessment techniques, should be an integral part of our instructional design, not just something added on for evaluation, i.e., determining grades.

Constructing the Test

6. ***Allow adequate time to construct essay questions.*** Although a five-question essay test can be constructed faster than a 50-item multiple-choice test, writing an effective essay question takes thought, and therefore time. One poorly designed essay question would have an effect similar to ten poor multiple-choice items.

7. ***Limit the use of essay questions to learning outcomes that cannot be satisfactorily measured by "objective" items.*** Given the serious limitations of essay tests, especially with respect to reliability, the recommendation is to use essay questions for assessment only when you have to. Especially, do not use essay questions to test facts, or learning at the lower levels of Bloom's taxonomy. (Bloom et al., 1956. For brief discussions of Bloom's taxonomy see Clegg and Cashin, 1986; and Gronlund, 1985b.)

8. ***Design the essay question to test only one or a few specific instructional objectives per question.*** This seems fairly clear; you must make explicit what you want to test (a necessary corollary is that you had to be clear about what you were trying to teach, i.e., expected the students to learn).

For example, the following is a poor essay question: "Why do animals migrate?" It might be better to ask: "Describe three hypotheses which might explain why animals migrate south in the fall of the year." This second version, however, points out that what is being tested basically is the students' memory of what was in the lecture or text—not a recommended use of essay questions. (See Clarence H. Nelson's chapter, "Evaluation in the Natural Sciences," in Dressel and Associates, 1961, for ways in which the students' understanding of these theories might be tested using "objective" items.)

A more appropriate example of a question testing a specific instructional objective, in this case a foreign language (Latin), is:

> Read the above passage and decide whether it was written by a classical or patristic Latin writer. Support your position by identifying and explaining specific phrases or passages which illustrate the characteristic writing style. Also identify phrases, etc., which might support the opposing position.

The objective was to assess, not simply students' passive understanding of the elements which characterize the two different styles, but also their ability to apply that knowledge in their reading. Of course, to do this the students also needed to have a certain proficiency in translating Latin. The instructor chose a passage unfamiliar to the students by an author whose writing contained elements of both styles (and a passage where the content did not serve as a clue), so it was possible for the students to make a case for either side. The primary point of the question was not whether the students correctly classified the author, but how good an argument they could make for their positions, and how aware they were of the contrary evidence. I consider this question to be at least at the Application level of Bloom's taxonomy.

9. ***Give preference to focused questions that can be answered briefly.*** When it fits your instructional objectives, several short essays will yield a more reliable score than fewer long questions. On the other hand, a short answer question is less likely to permit the students to demonstrate complex mental processes. Also, if an instructional objective can be tested by a short essay, perhaps it can also be tested by a multiple-choice item.

10. ***The question should clearly indicate the task(s)*** the students are to address with respect to both content and process. On one history of philosophy exam the students were given the following topic, "Locke: the key to Hume." While I applaud the creativity of the instructor, the question can be improved. For example:

> Locke: the key to Hume. Discuss the influence of the philosophy of Locke on Hume's theory of knowledge. **OR:**
>
> Locke: the key to Hume. Discuss the similarities and differences in the philosophies of John Locke and David Hume with respect to: the origin and relation ideas, the nature of belief, (etc.)

Gronlund (1985a, p. 220) provides a list of 12 types of thought questions and sample item stems, e.g.:

> Synthesizing: Describe a plan for . . .
> Evaluating: Describe the strengths and weaknesses . . .

Hopkins and Stanley (1981, pp. 214–216) list 21 types of essay questions, e.g.:

> Inferential thinking: Discuss whether the authors of this text are likely to use essay tests frequently in their measurement classes. Support your opinion with principles and recommendations given in the text.

One very helpful way to determine whether you have clearly specified the task is to give your essay questions to colleagues and see if they understand the questions. Also, ask them what instructional objective(s) they think you are trying to test with each question.

11. ***Make explicit the approximate time or length for each question, and/or the number of points.*** This is especially important if the questions are not weighted equally. Therefore, we might add to the "Locke" question above: (50 points, spend about 30 minutes, five pages on this question).

12. ***Provide sufficient time*** for the students to write the answer. See how long it takes you or a colleague to write an answer, then allow the student several times that amount of time. You do not want your tests basically to assess writing speed.

13. ***Use novel questions;*** otherwise you are testing memory. Novelty can provide interest, and therefore motivation, for the students. One psychology professor teaching a Systems of Psychology course asked the students to imagine that they were a rat in the lab of specific psychologists, and then describe what might happen to them with respect to a number of experimental variables.

14. ***Avoid optional questions,*** i.e., letting the students choose which question(s) they will answer. The only advantage is student morale, and the reasons against providing the students with a choice are persuasive:

A. ***The students are taking different tests.*** Thus, there is no common basis for comparison and the scoring becomes unreliable. It is almost impossible to write several essay questions which are of equal difficulty; the result is that different students are taking tests of varying difficulty but you will grade these the same. This may penalize the "better" students because they may choose the more difficult (challenging) questions and so will not score as well as students who choose the easier questions.

B. ***In real life we usually cannot pick our problems.*** In the world of work, at home, and in society, we are expected to address all of the major issues facing us, not the four out of five we feel most competent to handle.

Exception: There is one notable exception to the recommendation to avoid optional questions, and that is with extended response questions where you wish to test a skill. The most common example is assessing writing skills. Often students are given many topics and told to choose one to write on. The hope is that the list of topics will be broad enough to enable every student to find a topic he or she knows about. Thus, differences in the final essays will not reflect differences in their knowledge of the content, but will only reflect differences in writing skills. The same argument applies to a variety of other processes or skills: critical thinking, public speaking, artistic expression, etc. However, when your primary purpose is to test command of content, providing optional questions is not advised.

15. ***Do NOT give the students a short list of essay questions to prepare BEFORE the test.*** Although the intent in doing this is usually to help the students, the results are often undesirable. Such essay tests may simply test the students' ability to memorize someone else's thinking. If the list is short, it may encourage the students not to study all of the content. This can be exacerbated if the students know they will only have to answer, say, two out of four questions. In such cases they may simply omit studying two of the questions.

16. ***Prepare the students to take the test.*** Consider whether part or all of a class session might profitably be spent letting the students respond to a typical sample of your essay questions and then discussing what you look for when scoring them. Using a question from a previous year where you kept samples of "A" papers, "B" papers, etc., could be even more helpful to the students.

Scoring Essay Tests

The following recommendations are made to enhance the reliability and validity of scoring essay tests. The goal is to insure as much as possible that differences in students' essay scores reflect differences in their respective achievement, and nothing else.

17. ***Fit the scoring approach to the type of essay question.*** Two approaches are described in the literature: analytical (point-score) and global (holistic). (See Mehrens and Lehmann, 1984, pp. 114–116, for a longer discussion.)

Analytic (point-score) Method. This method is recommended for restricted-response questions. The ideal or model answer is broken down into several specific points regarding content. A specific subtotal point value is assigned to each. When reading the exam, you need to decide how much of each maximum subtotal you judge the student's answer to have earned.

Global (holistic) Method. This is recommended for extended-response questions. The rater reads the entire essay and makes an overall judgment about how successfully the student has covered everything that was expected in the answer and assigns the paper to a category (grade). Generally, five to nine categories are sufficient. Ideally, all of the essays should be read quickly and sorted into five to nine piles, then each pile reread to check that every essay has been accurately (fairly) assigned to that pile which will be given a specific score or letter grade.

18. When using analytical scoring for restricted-response questions, ***outline the model (ideal or acceptable) answer BEFORE you begin to read the essays.*** The specificity of the answer, however, may vary with the question. It is recommended that you read a sample of the actual essays before you begin to assign scores. Ideally, you should read all of the essays quickly to check (and perhaps modify) your model answer, then reread all of them to assign scores. In practice this often is not feasible, but remember that your goal is to have the students' scores reflect achievement on a common task. Also, you want to ***use realistic standards;*** reading several or all of the essays before assigning scores helps achieve this; having a colleague read your model answer would also help.

Ebel and Frisbie (1986, p. 127) suggest that some of the common reasons that students do not obtain maximum credit are:

(1) answer includes incorrect statements,
(2) relevant material is omitted,
(3) irrelevant material is included,
(4) student commits errors in logic, reaches unsound conclusion(s),

(5) student writes unclear answer, often because of poor writing skills (or poor handwriting), and
(6) student commits flagrant errors in grammar, spelling, etc.

Your model answer should provide guidelines for you to make an accurate assessment of at least the first three.

19. ***Keep the identity of the student anonymous.*** Your score should reflect your assessment of the adequacy of the answer and not anything else you know about the student.

20. ***Score one answer at a time*** (in a single, uninterrupted session if feasible). Experience suggests that most of us do not read all of the essays first before rereading them to assign a score. Reading the answers to single question (or mathematical problem) in a single sitting is at least a partial help for insuring that the students are being scored based upon a common set of criteria.

21. ***Shuffle the exams after scoring each question when the test consists of several items.*** This is suggested for a couple of reasons. First, if you have not followed some of the recommendations suggested above about clearly defining our model answer before assigning grades, there is a tendency to change expectations while reading the answers, e.g., you may become more and more depressed as you read the students' answers because no one has gotten it "right" and so tend to score essays read later more leniently. Second, if you keep the exams in the same order, you are more likely to be influenced by how well (or poorly) that student did on the previous question. Related to this, it is helpful to have the students start each answer on a new page so that you cannot see the score on the previous question.

22. ***Decide beforehand how you will handle grammar, spelling, handwriting, etc.*** Obviously, if you cannot read the student's writing, you cannot score his or her answer. If the writing (composition) is so unclear that you do not understand, you should not give credit. However, there are many instances where the substance of an answer is understandable, but there are obvious errors in spelling, grammar, etc. Sometimes students will take the position that these things should only influence the grade in English courses. However, college educated people should be able to write clear, grammatical, correctly spelled English. Even engineers and business people write reports and letters. You, however, must decide how much these things will count, especially in light of the time pressures typical in essay testing. You should also inform the students of your grading criteria well before they take any test.

23. ***When feasible, use multiple readings and/or readers.*** For really important essays, like undergraduate theses, two separate readings and scores are desirable, with the grade being the average of the two scores. Better yet, having two separate readers score the test is desirable. This is often done with essays used for placement in English composition, senior comprehensive exams, and the like. It is an effective, but time consuming, way to improve reliability.

24. ***Provide extensive comments.*** Although one of the purposes of an essay test, paper, etc., is to assess the students' past learning, they can and should also be used to help students continue to learn. Providing extensive comments, not just a grade, is a effective way to do this. One practice is to have the students write on every other page of the bluebook, leaving the opposite page for instructor commments. Since we can talk faster than we can write, some instructors, e.g., those teaching writing, have each student purchase an audiotape and the instructor records extensive comments on that tape. Because of time pressures comments are likely to be brief, but be sure that they are at least clear. Again, occasionally checking with a colleague can be very informative. If you use a code, tell the students; better yet, put it in writing. One student commented after an entire semester of receiving essays with check marks that he did not know whether the check meant good, or bad, or important, or what.

25. ***Consider keeping a test file on your essay questions.*** Over time you can develop a collection of essay questions (math problems, paper assignments, etc.) for specific instructional objectives. Your file should include your instructional objective(s), the question (problem) itself, any improvements that seem appropriate based upon past use, AND a record of how well the students performed on the question and any comments you have. Ideally, keep samples of "A" answers, "B" answers, etc. There is no sense in reinventing the wheel every year, especially if it is your own "wheel."

Conclusion

Despite serious limitations, especially with respect to reliability, essay tests have definite strengths, the most notable being their ability to test writing composition skills. Although most of the recommendations in this paper focus on the essay test per se, they can also be applied to papers, project reports, mathematical problems, artistic productions, and the like, with the necessary adaptations. Also, the focus of this paper has been on using essays to assess student learning. However, most of the caveats expressed are of less concern when using writing essays as a learning experience to provide students with feedback about their performance. It is hoped that the recommendations offered in this paper will help college teachers improve their essay tests.

References and Further Readings

The four references that are followed by an asterisk are standard texts on educational measurement. They contain chapters on essay tests (as well as "objective" tests, grading, etc.). I would take the position that every college teacher should have an educational measurement text in his or her library. There are many available—these are four books particularly worth your consideration.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). ***Taxonomy of education objectives: Handbook I, the cognitive domain.*** New York: David McKay.

Clegg, V. L., & Cashin, W. E. (1986). ***Improving multiple-choice tests.*** (IDEA Paper No. 16). Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

Coffman, W. E. (1971). Essay examinations. In R. L. Thorndike (Ed.) ***Educational measurement*** (2nd ed., pp. 271–302). Washington, D.C.: American Council on Education.

Dressel, P. E., & Associates. (1961). ***Evaluation in higher education.*** Boston: Houghton Mifflin.
Despite its 1961 publication data, this is one of a few books that gives extensive treatment to testing approaches for several different academic fields at the college level, e.g., natural sciences, social sciences, humanities.

Ebel, R. L., & Frisbie, D. A. (1986). ***Essentials of educational measurement*** (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.*

Gronlund, N. E. (1985a). ***Measurement and evaluation in teaching*** (5th ed.). New York: Macmillan.*

Gronlund, E. E. (1985b). ***Stating objectives for classroom instruction*** (3rd ed.). New York: Macmillan.

Hopkins, K. D., & Stanley, J. C. (1981). ***Educational and psychological measurement and evaluation*** (6th ed.). Englewood Cliffs, NJ: Prentice-Hall.*

Mehrens, W. A., & Lehmann, I. J. (1984). ***Measurement and evaluation in education and psychology*** (3rd ed.) New York: Holt, Rinehart and Winston.*

Center for Faculty Evaluation and Development
Kansas State University
1615 Anderson Avenue
Manhattan, KS 66502–4073
1-800-255-2757 or (913) 532-5970
or FAX (913) 532-5637