CHAPTER

10

MEASURING COMPLEX ACHIEVEMENT: ESSAY QUESTIONS

Objectives

From this chapter you should be able to:

1. Describe the uses of essay questions.
2. Describe the advantages and limitations of essay questions.
3. Distinguish between restricted-response and extended-response essays.
4. Construct essay questions.
5. Construct scoring rubrics for essays.

Some important learning outcomes may best be measured by the use of open-ended essay questions or other types of performance assessments. Essay questions provide the freedom of response that is needed to adequately assess the ability of students to formulate problems; organize, integrate, and evaluate ideas and information; and apply knowledge and skills.

Up to this point, our main concern has been with objective test items. We noted that such items can measure a variety of learning outcomes, from simple to complex, and that the interpretive exercise is especially useful for measuring complex achievement. Despite this wide applicability of objective-item types, there remain significant instructional outcomes for which no satisfactory objective measurements have been devised. These include such outcomes as the ability to recall, organize, and integrate ideas; the ability to express oneself in writing; and the ability to create rather than merely identify interpretations and applications of data. Such outcomes require less structuring of responses than objective test items, and it is in the measurement of these outcomes that written essays and other performance-based assessments are of greatest value.

In this chapter, we consider the most familiar form of performance-based assessment: the essay question. Other types of performance-based assessments (which include gathering information, making oral presentations, conducting experiments, repairing or

manipulating equipment, and so on) are considered in Chapter 11. Purposeful collections of student work into portfolios, which may include a wide variety of different types of assessments (e.g., written essays and other types of performance assessments), are considered in Chapter 12. Teacher observations, peer appraisals, and self-reports are considered in Chapter 13.

FORMS AND USES OF ESSAY QUESTIONS

We focus our discussion of the essay question on its use in the measurement of complex achievement. We recognize, however, that many teachers use essay questions to measure knowledge of factual information. It certainly can be useful to ask students to generate, in their own words, the plot of a story, the causes of a historical event, or the steps in a scientific process, all of which may be provided by a text. Although measuring such knowledge of factual information with essay questions is useful and valid, it does not tap the full potential of essay questions.

The distinctive feature of essay questions is the freedom of response. Students are free to construct, relate, and present ideas in their own words. Although this freedom enhances the value of essay questions as a measure of complex achievement, it introduces scoring difficulties that make essays inefficient as a measure of factual knowledge. For most purposes, knowledge of factual information can be more efficiently measured by some type of objective item. Essay questions should be used primarily to measure those learning outcomes that are not readily measured by objective test items. The special features of essay questions can be utilized most fully when their shortcomings are offset by the need for such measurement. Learning outcomes concerned with the abilities to conceptualize, construct, organize, integrate, relate, and evaluate ideas require the freedom of response and the originality provided by essay questions and other performance assessments. In addition, these outcomes are of such great educational significance that the expenditure of energy in the difficult and time-consuming task of evaluating the answers can be easily justified.

Essay tests and other performance-based assessments can also be justified on the grounds that the performances required correspond more closely to the larger instructional goals and objectives than discrete factual-knowledge questions. Indeed, the validity of measurement of complex achievement may be enhanced by the use of essay tests and other performance-based assessments. Furthermore, tests send a message of what is important to learn and be able to do. Just consider how frequently teachers are asked the question, "Will this be on the test?" The form of the assessment provides a model. Thus, it is often argued that if you want students to be able to communicate in writing, then they not only need to be encouraged to write but also have to be required to do so when it counts.

As implied by the previous comments, essay assessments can be useful ways of assessing student understanding and ability to organize and apply information in a content area such as history, civics, literature, science, or mathematics. In any of these or other content areas, the essay assessment allows teachers to evaluate how well students can communicate ideas. Essay assessments are, of course, also widely used where the main focus is on evaluating student writing without regard to any particular subject matter

content. In the latter case, the emphasis is more likely to be on the form of the writing distinguishing, for example, between narrative essays, expository essays, and persuasive essays. Essay assessments may also be used to focus teacher and student attention on the writing process itself through the use of various prewriting activities (e.g., discussion, listing and organizing ideas, constructing outlines, and clarification of audience) as well as the initial drafting and revision of essays.

The freedom of response provided by essay questions is not an all-or-nothing affair but, rather, a matter of degree. At one extreme, the response is almost as restricted as that in the short-answer objective item, in which a sentence or two may be all that is required. At the other extreme, students are given almost complete freedom in constructing their responses. The written essay may be several pages in length. Where the emphasis is on the writing process itself, the essay responses may include prewriting responses such as notes, lists of ideas, and outlines as well as initial drafts and revisions. Although variations in freedom of response tend to fall along a continuum between these extremes, essay questions can be conveniently classified into two types: restricted-response questions and extended-response questions or assignments.

Restricted-Response Essay Questions

The restricted-response question usually limits both the content and the response. The content is usually restricted by the scope of the topic to be discussed. Limitations on the form of response are generally indicated in the question:

EXAMPLES Describe two situations that demonstrate the application of the law of supply and demand. Do not use those examples discussed in class.

State the main differences between the Vietnam War and previous wars in which the United States has participated.

Why is the barometer one of the most useful instruments for forecasting weather? Answer in a brief paragraph.

Write the verbal instructions you would give to a friend on the telephone so that the friend could draw a triangle on a piece of graph paper with sides that have relative lengths of 3, 4, and 5 units.

What is measured on an essay such as the one asking students to state the differences between the Vietnam War and previous wars depends on a student's previous instructional experiences. If the textbook or recent class presentations have explicitly discussed ways in which the Vietnam War was different from previous wars, then the student's task is simply to demonstrate an understanding of this material and to put it in his or her own words. That is, the essay question is simply a measure of comprehension. If the essay question presents the student with his or her first opportunity to think about the Vietnam War in terms of differences from previous wars, however, then the essay requires analysis and higher-level thinking.

Another way of restricting responses in essay questions is to base the questions on specific problems. For this purpose, introductory material like that used in interpretive exercises can be presented. Such items differ from objective interpretive exercises only by the fact that essay questions are used instead of multiple-choice or true–false items.

EXAMPLE There is a broad consensus among medical scientists that smoking is damaging to the health of both smokers and those who are exposed to cigarette smoke on a regular basis. Some cities have passed laws banning smoking inside all public buildings. Some people have argued against such regulations on the grounds that smoking bans violate the freedom of choice of individual smokers.

(A) Indicate whether you agree or disagree with the underlined part of the last statement.
(B) Support your position.

Because the restricted-response question is more structured than the extended-response essay considered next, it is most useful for measuring learning outcomes requiring the interpretation and application of data in a specific area. In fact, any of the learning outcomes measured by an objective interpretive exercise also can be measured by a restricted-response essay question. The difference is that the interpretive exercise requires students to select the answer, whereas the restricted-response question requires them to supply it. In some instances, the objective interpretive exercise is favored because of the ease and reliability of scoring. In other situations, the restricted-response essay question is better because of its more direct relevance to the learning outcome (e.g., the ability to formulate valid conclusions).

Although restricting students' responses to essay questions makes it possible to measure more specific learning outcomes, these same restrictions make them less valuable as a measure of those learning outcomes emphasizing integration, organization, and originality. Restricting the scope of the topic to be discussed and indicating the nature of the desired response limit the student's opportunity to demonstrate these behaviors. For higher-order learning outcomes, greater freedom of response is needed.

Extended-Response Essays

The extended-response question or assignment allows students to select any factual information that they think is pertinent, to organize the answer in accordance with their best judgment, and to integrate and evaluate ideas as they deem appropriate. This freedom enables them to demonstrate their ability to analyze problems, organize their ideas, describe in their own words, and/or develop a coherent argument. If analysis, organization, integration, creative expression, and evaluation skills are emphasized in the grading of the essays as well as in instruction, this form of assessment also makes clear the value that is placed on these higher-order skills. On the other hand, this same freedom that enables the demonstration of creative expression and other higher-order skills makes the extended-response question inefficient for measuring more specific learning outcomes and introduces scoring difficulties:

EXAMPLES Imagine that you and a friend found a magic wand. Write a story about an adventure that you and your friend had with the magic wand.

Compare developments in international relations in the administrations of President William Clinton and President George W. Bush. Cite examples when possible.

Evaluate the significance of the sea captain's pursuit of the white whale in *Moby Dick*.
Describe the influence of Mendel's laws of heredity on the development of biology as a science.
Write a scientific evaluation of the Copernican theory of the solar system. Include scientific observations that support your statements.

The need to measure a student's global attack on a problem can be easily defended. The thinking and problem-solving skills measured by objective interpretive exercises and restricted-response essay questions seldom function in isolation. In a natural situation they operate together in a manner that includes more than a sum of the skills involved. These skills interact with one another and with the knowledge and understanding the problem requires. Thus, it is not just the skills we are measuring but also how they function together.

Both teachers and test specialists agree that the extended-response question does require complex behaviors that cannot be measured by more objective means. But they often differ in their level of concern about the difficulty of scoring extended written responses in a way that can satisfactorily measure these behaviors. Test specialists point out that unless considerable attention is given to the choice of questions and to scoring procedures, the scoring may be too unreliable to yield defensible measurement. Nevertheless, many teachers continue to use the extended-response question to measure student achievement without adequate attention to the complexities involved in the construction and scoring of such questions. Neither a hard-line measurement position that rejects extended essays as an approach to measurement nor one that ignores the difficulties of scoring seems to contribute much to the valid measurement of student achievement. It seems more sensible to identify the complex skills we want to measure, formulate questions that elicit these skills, evaluate the results as reliably as we can, and then use these admittedly limited data as the best evidence we have available.

SUMMARY COMPARISON OF LEARNING OUTCOMES MEASURED

The restricted-response essay question can measure a variety of complex learning outcomes similar to those measured by the objective interpretive exercise. The main difference is that the interpretive exercise requires students to select the answer, and the restricted-response question requires the student to supply the answer. In comparison, extended-response essay assessments measure more general learning outcomes, such as the abilities to organize, integrate, evaluate, and express ideas. They may be used to measure writing skills as well as the understanding and ability to apply subject-matter content knowledge. A comparison of the types of complex learning outcomes measured by each of these types of assessment is presented in Table 10.1. The learning outcomes in the table, of course, merely suggest the types of learning outcomes that may be measured. With slight modifications, an infinite variety of outcomes can be stated in each area. The freedom of response to essay questions is a matter of degree, and thus the functions of the restricted-response question and the extended-response question often overlap.

Table 10.1
Types of complex learning outcomes measured by essay questions and objective interpretive exercises

Type of Assessment Item	*Examples of Complex Learning Outcomes That Can Be Measured*
Objective interpretive exercises	Ability to— • identify cause-and-effect relationships • identify the application of principles • identify the relevance of arguments • identify tenable hypotheses • identify valid conclusions • identify unstated assumptions • identify the limitations of data • identify the adequacy of procedures (and similar outcomes based on the pupil's ability to *select* the answer)
Restricted-response essay questions	Ability to— • explain cause-and-effect relationships • describe applications of principles • present relevant arguments • formulate tenable hypotheses • formulate valid conclusions • state necessary assumptions • describe the limitations of data • explain methods and procedures (and similar outcomes based on the pupil's ability to *supply* the answer)
Extended-response essays	Ability to— • produce, organize, and express ideas • integrate learnings in different areas • create original forms (e.g., designing an experiment) • summarize (e.g., writing a summary of a story) • construct creative stories (e.g., narrative essays) • explain concepts or principles (e.g., expository essay) • persuade a reader (e.g., persuasive essay) (and similar outcomes based on a pupil's ability to *write* an essay for a given purpose)

ADVANTAGES AND LIMITATIONS OF ESSAY QUESTIONS

Advantages

A major advantage of the essay question is that it measures complex learning outcomes that cannot be measured by other means. But the use of essay questions does not guarantee the measurement of complex achievement. To do so, essay questions must be as carefully constructed as objective test items. The course objectives pertinent to complex achievement must be defined in terms of specific learning outcomes, and

the essay questions must be phrased in a way that will require students to engage in the targeted thinking skills. When a table of specifications is used in planning for the assessment, it is simply a matter of constructing the questions in accordance with the specifications.

A second advantage of the extended-response essay is its emphasis on the integration and application of thinking and problem-solving skills. Although objective items such as the interpretive exercise can be designed to measure various aspects of complex achievement, the ability to integrate and apply these skills in a general attack on a problem is best measured by extended-response essay questions.

Perhaps the most obvious advantage of essay assessments is that they enable the direct evaluation of writing skills. In some instances, the evaluation of specific writing skills may be combined with the assessment of subject-matter knowledge and understandings (e.g., communication of mathematical or scientific principles, ideas, and concepts). In other cases, the assessment of writing skills may be the sole or primary purpose (e.g., skill in developing characters in a narrative story or writing mechanics).

Another commonly cited advantage of the essay question is its ease of construction. This factor has led to the widespread use of essay questions by classroom teachers. In a matter of minutes, most teachers can formulate several essay questions, an attractive feature for the busy teacher. This apparent advantage can be very misleading, however. Constructing essay questions that require the conceptual understanding and thinking skills emphasized in a particular set of learning outcomes takes considerable thought and effort. When ease of construction is stressed, it usually refers to the common practice of dashing off questions with little regard for the course objectives. In such cases, there is some question whether ease of construction can be considered an advantage. In addition to the invalidity of the measurement, evaluating the answers to carelessly developed questions tends to be confusing. Moreover, valid scoring of responses to any essay question requires great care in the development and application of scoring rubrics, and providing written comments and suggestions on student essays that can help students improve their writing is both highly desirable and time consuming.

Finally, the potentially most important advantage of the essay question is its contribution to student learning. The contribution to learning can be direct. The process of preparing a response to an extended-response essay question, for example, may also be an effective learning exercise. The effects on learning can also be indirect. The models of what students are expected to do in response to essay questions often coincide with and encourage effective learning activities.

Limitations

The most commonly cited limitation of the essay question is the unreliability of the scoring. Over the years, various studies have shown that written essays are scored differently by different teachers and that even the same teachers score responses differently at different times. The poor reliability across scorers, however, is frequently the result of failure to identify clearly the learning outcomes being measured and the failure to establish well-defined scoring rubrics.

Evaluating essays without adequate attention to the learning outcomes being measured and the scoring rubrics to be used is like "three blind men appraising an elephant." One

teacher stresses factual content; one, organization of ideas; and another, writing skill. With each teacher evaluating the degree to which different learning outcomes are achieved, it is not surprising that scoring diverges. Even variations in scoring by the same teacher can probably be explained to a large extent by inadequate attention to learning outcomes and scoring rubrics. When the evaluation of answers is not guided by clearly defined outcomes and scoring rubrics, it tends to be based on less stable, intuitive judgments. Although the judgmental scoring of essay responses will always have some degree of unreliability, scoring reliability can be greatly increased by clearly defining the outcomes to be measured, properly framing the questions, carefully following scoring rules, and obtaining practice in scoring.

A closely related limitation of essay questions is the amount of time required for scoring the responses. If the scoring is done conscientiously and helpful feedback is provided to students, even a small number of papers may require several hours of scoring time. If the classes are large and several extended-response essay questions are used, conscientious scoring becomes practically impossible. Ironically, most of the suggestions for improving the scoring of responses to essay questions require more time, not less, as might be hoped. The only practical solution is to reserve the use of extended-response essay questions for those learning outcomes that cannot be measured well objectively. With fewer essay questions to score in a given test, more time will be available for evaluating the answers.

Another shortcoming of essay questions is the limited sampling of content they provide. So few questions can be included in a given test that some areas are measured thoroughly while many others are neglected. This inadequate sampling makes essay questions especially inefficient for measuring knowledge of factual information. For such outcomes, we can use objective test items and reserve essay questions, especially extended-response questions, for measuring complex achievement. This does not eliminate the sampling problem, however, because we would also like an adequate sample of complex behaviors. When we use essay questions, we should try to obtain as representative a sample of learning outcomes as possible. One way of doing this is to accumulate evidence from a series of essay questions administered at different times throughout the school year. The collection of the results throughout the year into portfolios of work, as is described in Chapter 12, can serve other important evaluation and communication functions.

SUGGESTIONS FOR CONSTRUCTING ESSAY QUESTIONS

The improvement of the essay question as a measure of complex learning outcomes requires attention to two problems: (a) how to construct essay questions that call forth the desired student responses and (b) how to score the answers so that achievement is reliably measured. Here we suggest ways of constructing essay questions, and in the next section we suggest ways of improving scoring, although these two procedures are interrelated.

1. **Restrict the use of essay questions to those learning outcomes that cannot be measured satisfactorily by objective items.** Other things being equal, objective measures have the advantage of efficiency and reliability. But when objective items are inadequate for measuring the learning outcomes, the use of essay questions can be easily defended

despite their limitations. Some of the complex learning outcomes, such as those pertaining to the organization, integration, and expression of ideas, will be neglected unless essay questions are used. By restricting the use of essay questions to these areas, the evaluation of student achievement can be most fully realized.

2. Construct questions that will call forth the skills specified in the learning standards. Like objective items, essay questions should measure the achievement of clearly defined content standards or instructional outcomes. If the ability to apply principles is being measured, for example, the questions should be phrased in such a manner that they require students to display their conceptual understanding or a particular skill. Essay questions should never be hurriedly constructed in the hope that they will measure broad, important (but unidentified) educational goals. Each essay question should be carefully designed to require students to demonstrate achievement defined in the desired learning outcomes. See the box "Some Types of Thought Questions and Sample Item Stems" for examples of the many types of questions that might be asked; the phrasing of any particular question will vary somewhat from one subject to another.

Some Types of Thought Questions and Sample Item Stems

Comparing

- Describe the similarities and differences between . . .
- Compare the following two methods for . . .

Relating Cause and Effect

- What are major causes of . . . ?
- What would be the most likely effects of . . . ?

Justifying

- Which of the following alternatives would you favor, and why?
- Explain why you agree or disagree with the following statement.

Summarizing

- State the main points included in . . .
- Briefly summarize the contents of . . .

Generalizing

- Formulate several valid generalizations from the following data.
- State a set of principles that can explain the following events.

Inferring

- In light of the facts presented, what is most likely to happen when . . . ?
- How would Senator X be likely to react to the following issue?

Explaining

- Why did the candle go out shortly after it was covered by the jar?
- Explain what President Truman meant when he said, "If you can't stand the heat, get out of the kitchen."

Persuading

- Write a letter to the principal to get approval for a class field trip to the state capital.
- Why should the student newspaper be allowed to decide what should be printed without prior approval from teachers?

Classifying

Group the following items according to . . .
What do the following items have in common?

Creating

List as many ways as you can think of for . . .
Make up a story describing what would happen if . . .

Applying

Using the principle of . . . as a guide, describe how you would solve the following problem situation.
Describe a situation that illustrates the principle of . . .

Analyzing

Describe the reasoning errors in the following paragraph.
List and describe the main characteristics of . . .

Synthesizing

Describe a plan for proving that . . .
Write a well-organized report that shows . . .

Evaluating

Describe the strengths and weaknesses of . . .
Using the given criteria, write an evaluation of . . .

Constructing essay questions in accordance with particular learning outcomes is much easier with restricted-response questions than with extended-response questions. The restricted scope of the topic and the type of response expected make it possible to relate a restricted-response question directly to one or more of the outcomes. The extreme freedom of the extended-response question makes it difficult to present questions so that the student's responses will reflect the particular learning outcomes desired. This difficulty can be partially overcome by indicating the basis on which the answer will be evaluated:

EXAMPLE Write a two-page statement defending the importance of conserving our natural resources. (Your answer will be evaluated in terms of its organization, its comprehensiveness, and the relevance of the arguments presented.)

Informing students that they should pay special attention to organization, comprehensiveness, and relevance of arguments defines the task, makes the scoring criteria explicit, and makes it possible to key the question to a particular set of learning outcomes. These directions alone will not, of course, ensure that the appropriate behaviors will be exhibited. It is only when the students have been taught the relevant skills and how to integrate them that such directions will serve their intended purpose.

3. Phrase the question so that the student's task is clearly indicated. The purpose a teacher had in mind when developing the question may not be conveyed to the student if the question contains ambiguous phrasing. Students interpret the question differently and give a hodgepodge of responses. Because it is impossible to determine which of the incorrect or off-target responses are due to misinterpretation and which to lack of achievement, the results are worse than worthless. They may actually be harmful if used to measure student progress toward instructional objectives.

One way to clarify the question is to make it as specific as possible. For the restricted-response question, this means rewriting it until the desired response is clearly defined.

EXAMPLE

Poor: Why do birds migrate?

Better: State three hypotheses that might explain why birds migrate south in the fall. Indicate the most probable one and give reasons for your selection.

The improved version presents the students with a definite task. Although some students may not be able to give the correct answer, they all will certainly know the type of response expected. Note also how easy it would be to relate such an item to a specific learning outcome, such as "the ability to formulate and defend tenable hypotheses."

When an extended-response question is desired, some limitation of the task may be possible, but care must be taken not to destroy the function of the question. If the question becomes too narrow, it will be less effective as a measure of the ability to select, organize, and integrate ideas and information. The best procedure for clarifying the extended-response question seems to be to give the student explicit directions concerning the type of response desired:

EXAMPLE

Poor: Compare the Democratic and Republican parties.

Better: Compare the current policies of the Democratic and Republican parties with regard to the role of government in private business. Support your statements with examples when possible. (Your answer should be confined to two pages. It will be evaluated in terms of the appropriateness of the facts and examples presented and the skill with which it is organized.)

The first version of the example offers no common basis for responding and, consequently, no frame of reference for evaluating the response. If students interpret the question differently, their responses will be organized differently, because organization is partly a function of the content being organized. Also, some students will narrow the problem before responding, thus giving themselves a much easier task than students who attempt to treat the broader aspects of the problem.

The improved version gives students a clearly defined task without destroying their freedom to respond in original ways. This is achieved both by specifying the scope of the question and by including directions concerning the type of response desired. See the box "The Importance of Writing Skill."

4. Indicate an approximate time limit for each question. Too often, essay questions place a premium on speed because inadequate attention is paid to reasonable time limits during the test's construction. As each question is constructed, the teacher should estimate the approximate time needed for a satisfactory response. In allotting response time, keep the slower students in mind. Most errors in allotting time needed are in giving too little time. It is better to use fewer questions and give more generous time limits than to put some students at a disadvantage.

The time limits allotted to each question should be indicated to the students so that they can pace their responses to each question and not be caught at the end of the testing time with "just one more question to go." If the assessment contains both objective and

essay questions, the students should, of course, be told approximately how much time to spend on each part of the test. This may be done orally or included on the test form itself. In either case, care must be taken not to create overconcern about time. The adequacy of the time limits might very well be emphasized in the introductory remarks so as to allay any anxiety that might arise.

5. Avoid the use of optional questions. A fairly common practice when using essay questions is to give students more questions than they are expected to perform and then permit them to select a given number. For example, the teacher may include six essay questions in a test and direct the students to respond to any three of them. This practice is generally favored by students because they can select those questions they know most about. Except for the desirable effect on student morale, however, there is little to recommend the use of optional questions. If students answer different questions, it is obvious that they are taking different tests, and so the common basis for evaluating their achievement is lost. Each student is demonstrating the achievement of different learning outcomes. As noted earlier, even the ability to organize cannot be measured adequately without a common set of responses because organization is partly a function of the content being organized.

The use of optional questions might also influence the validity of the test results in another way. When students anticipate the use of optional questions, they can prepare responses on several topics in advance, commit them to memory, and then select questions to which the responses are most appropriate. During such advance preparation, it is also possible for them to get help in selecting and organizing their response. Needless to say, this provides a distorted measure of the student's achievement, and it also tends to have an undesirable influence on study habits, as intensive preparation in a relatively few areas is encouraged.

Of course, there are learning outcomes that involve in-depth study of topics that are shaped and defined by students. Evaluation of student work on topics of their own choosing is important for such learning outcomes. The assessment of such outcomes, however, is better approached through the assignment of projects than by an essay test. See the "Checklist" box to evaluate essay questions you construct.

The Importance of Writing Skill

Performance on an essay test depends largely on writing ability. If students are to be able to demonstrate the achievement of higher-level learning outcomes, they must be taught the thinking and writing skills needed to express themselves. This means teaching them how to select relevant ideas, how to compare and relate ideas, how to organize ideas, how to apply ideas, how to infer, how to analyze, how to evaluate, and how to write a well-constructed response that includes these elements. Asking students to "compare," "interpret," or "apply" has little meaning unless they have been taught how to do these things. This calls for direct teaching and practice in writing, in an atmosphere that is less stressful than an examination period. Use of analytic scoring criteria that give separate scores for characteristics such as the quality of ideas, use of examples, and use of supporting evidence and ones dealing with writing mechanics such as grammar, punctuation, and spelling can improve scoring and, if communicated to students, can both guide their efforts in constructing essays and lead to improvements of specific writing skills.

CHECKLIST

Reviewing Essay Questions

	Yes	No
1. Is this the most appropriate type of task to use?	___	___
2. Are the questions designed to measure higher-level learning outcomes?	___	___
3. Are the questions relevant to the intended learning outcomes?	___	___
4. Does each question clearly indicate the response expected?	___	___
5. Are students told the bases on which their answers will be evaluated?	___	___
6. Are generous time limits provided for responding to the questions?	___	___
7. Are students told the time limits and/or point values for each question?	___	___
8. Are all students required to respond to the same questions?	___	___
9. If revised, are the questions still relevant to the intended learning outcomes?	___	___
10. Have the questions been set aside for a time before reviewing them?	___	___

SCORING CRITERIA

Clear specification of scoring criteria in advance of administering essay questions can contribute to improved reliability and validity of the assessment. Planning how responses will be scored will frequently lead to rethinking and clarification of the questions so that students have a clearer idea of what is expected. Informing students of the scoring criteria that will be used in evaluating their responses also can enhance the validity of the assessments because students are more likely to focus their efforts in the direction intended by the teacher.

After the assessment has been administered, it is often useful to do an initial review of the responses to a single question. Based on the initial review, a few exemplar or "anchor" responses may be identified that most clearly correspond to the levels of the scoring rubric. The comparability and fairness of scores assigned to student responses can be enhanced by comparing each response to the selected anchor responses.

It is important that scores or levels identified in a scoring rubric be descriptive and not merely judgmental in nature. It is better, for example, to define a level of the rubric as "writing is clear and thoughts are complete" than to only characterize the level as "excellent." Reliability, comparability, and fairness of scores are enhanced by clear descriptions.

Scoring Rubrics for Restricted-Response Essay Questions

In many instances, scoring guides for restricted-response essay questions are most readily constructed starting with the teacher writing an example of an expected response. If the student is asked to describe three factors that contributed to the start of the Civil War, for example, the teacher might construct a list of acceptable reasons and simply give the student 1 point for each of up to three reasons given from the list. In the example given earlier where students are asked to write a paragraph explaining why a barometer is one

of the most useful instruments in forecasting weather, the teacher might list key ideas that would need to be there for the student to get full credit as well as the level of explanation that would be awarded partial credit.

Analytic Scoring Rubrics for Extended-Response Essays

Analytic scoring rubrics enable a teacher to focus on one characteristic of a response at a time. The separation of characteristics such as writing mechanics from the quality of the content of the essay can be especially useful. Separate scores for characteristics such as these provide the student with clearer feedback about the strengths and weaknesses of the response.

Analytic scores for writing skills may consist of just two broad categories such as rhetorical effectiveness and conventions or content quality and mechanics. Sometimes finer distinctions are useful. The scoring rubrics used by the state of Oregon for its statewide writing assessment consists of seven analytic dimensions:

1. Ideas and Content
2. Organization
3. Voice
4. Word Choice
5. Sentence Fluency
6. Conventions
7. Citing Sources (For use on classroom assignments requiring research)

Scoring rubrics for 5-point and 6-point ratings are available on-line at the Northwest Regional Educational Laboratory (NWREL) website at *http://educationnorthwest.org/resource/464*. The NWREL analytic scoring rubrics are presented for seven dimensions or "traits." The first NWREL dimension is Ideas and the seventh dimension is Presentation. The remaining five dimensions are the same as the Oregon assessment. The specification of a score of 3 on the Organization dimension is shown in the box showing a sample scoring rubric for the beginning writer (K–2). Similar descriptions are given for score points of 1, 2, 4, and 5 for this and the other dimensions.

These lists, together with the actual descriptions of rubrics, may provide a useful starting point for constructing analytic scoring dimensions for use in the classroom. For any such list, decisions would need to be made about the number of score points to use and the criteria for determining the score level on each dimension. Scoring rubrics such as the one available on-line from the NWREL illustrate ways in which the individual score points can be described.

Another example illustrating descriptions of score points on analytic dimensions is shown in Table 10.2. The examples in the table were adapted from work by Gearhart, Herman, Baker, and Whittaker (1994). Six scale points on four analytic scales and an overall general impression dimension are described. Scoring rubrics such as these are useful in scoring expository essays or descriptive summaries. Variations may be useful for other types of essays. For example, in scoring a persuasive essay, additional dimensions for rating the use of supporting evidence, distinguishing between fact and opinion, and determining the coherence of the argument may be desirable for giving students feedback on how to make their argument more effective.

Table 10.2
Example analytic scales for expository essays or descriptive summaries

Score	*General Impression*	*Focus/ Organization*	*Language*	*Elaboration*	*Mechanics*
6	Exceptional achievement	• Clearly stated main idea • Unified focus and organization • Effectively orients reader	• Specific and concrete • Details consistent with intent • Details create clear, vivid image	• Extended elaboration of one main point	• One or two minor errors • No major errors
5	Commendable achievement	• Stated or implied main idea • Focused and organized • Effectively orients reader	• Specific sensory details • Most details consistent with intent	• Full elaboration of one main point	• A few minor errors • No more than one major error
4	Adequate achievement	• Main idea present but may not maintain consistent focus • Some orientation of reader	• Some specific details • Details usually clear • Generally clear images	• Moderate elaboration of main point	• Some minor errors • One or two major errors • Errors do not cause reader confusion
3	Some evidence of achievement	• Main idea not clear • Usually on topic, but with some digressions	• Few or inconsistent details • Some details, but all may not be appropriate	• Restricted elaboration of main point	• Some minor and some major errors • Some cause reader confusion
2	Limited evidence of achievement	• Vague indication of main idea or focus • Significant digressions • No sense of closure	• Little concrete language • Simple or generic naming	• Limited elaboration of main point	• Many minor and major errors • Errors interfere with reader understanding
1	Minimal evidence of achievement	• No apparent main idea • No apparent plan or coherence	• No concrete language	• No elaboration of any point or central statement	• Many major errors causing reader confusion

Source: Adapted from Gearhart, Herman, Baker, and Whittaker (1994).

Example of the NWREL Scoring Rubric to Be Considered a Developing Writer (3 on a 5-point scale) on the Organization Dimension for the Beginning Writer (K–2)

- "Beginning and middle are present, but no ending."
- "Transitions rely on connective 'and'."
- "Sequencing is adequate."
- "Pacing is adequate."
- "Simple title (if required) works."
- "Structure is present and works."

Narrative essays rubrics for five analytic dimensions were used by primary classroom teachers in several studies (see Wolf & Gearhart, 1997). The five dimensions are the following:

1. Theme, including considerations of degree to which it is explicit or implicit and the degree to which it is didactic or revealing
2. Character, including the degree to which the characters are flat and static or "round" and dynamic
3. Setting, including the degree to which the setting is simple or multifunctional and the degree to which it is merely part of the backdrop or essential to the story
4. Plot, including the degree to which the plot is simple or complex and the degree to which it is static or presents conflict
5. Communication, including the degree to which the story is context based or reader considerate and the degree to which it is literal or symbolic

For each of these dimensions, descriptions of six levels of performance are described (see Wolf & Gearhart, 1997 [available at *http://www.cse.ucla.edu/products/reports.php*]; or see the home page for the Center for Research on Evaluation, Standards, and Student Testing [CRESST] at *http://www.cse.ucla.edu*).

Holistic Scoring Rubrics for Extended-Response Essays

As the name suggests, holistic scoring rubrics yield a single overall score taking into account the entire response. Holistic scoring rubrics can generally be constructed more rapidly, and they generally can be used to score a set of essay responses more rapidly than analytic scoring rubrics. These advantages must be weighed against the major disadvantage that they do not provide students with feedback on specific aspects of the response that are strong and ones where improvement is needed. Of course, such feedback can be provided by marginal notes and comments that the teacher writes on the student's paper, but holistic scores alone provide less specific guidance to the student than analytic scores. It is also the case that the ease of construction of a set of labels (e.g., excellent, good, adequate, promising but has major shortcomings, weak, and inadequate) is no real advance of the traditional A, B, C, D, and F marks and provides little if any real guidance to the teacher in scoring or to the student in understanding what is expected. Such labels alone fall short of what is meant by a scoring rubric.

A holistic scoring rubric, like an analytic scoring rubric, needs to have the scores or labels elaborated by statements of the characteristics of the response that deserve the score of "excellent" or "promising but has major shortcomings." The National Assessment

Table 10.3
NAEP holistic scoring rubric for writing

Score	Description of Score Point
1	"*Response to topic* with little information pertinent to task."
2	"*Undeveloped response to the task* in which students began to respond, but did so in a very abbreviated, confusing, or disjointed manner."
3	"*Minimally developed:* a response in which student provided a response to the task that was brief, vague, and somewhat confusing."
4	"*Developed:* a response to the task that contained the necessary elements, but may have been unevenly developed or unelaborated."
5	"*Elaborated:* a well developed and detailed response that may have gone beyond the essential elements of the task."
6	"*Extensively elaborated:* a response that shows a high degree of control over the various elements of writing. Compared with papers given a rating of '5,' those rated '6' may have been similar in content, but they were better organized, more clearly written, and less flawed."

Source: Applebee et al. (1994, p. 204).

of Educational Progress (NAEP) writing assessment uses a 6-point holistic scoring rubric shown in Table 10.3.

SUGGESTIONS FOR SCORING ESSAY QUESTIONS

Improving the reliability of scoring essay questions begins long before the questions are administered. The first step is to decide what learning outcomes are to be measured. This is followed by phrasing the questions and the scoring rubrics in accordance with the learning outcomes and including explicit directions concerning the type of answers desired. Only when both the students and the teacher understand the task to be performed can reliable scoring be expected. No degree of proficiency in evaluating answers can compensate for poorly designed and phrased questions.

When the necessary preliminary steps have been taken in constructing essay questions, the following suggestions can be used effectively to increase the reliability of the scoring:

1. Prepare an outline of the expected answer in advance. This should contain the major points to be included, the characteristics of the answer (e.g., organization) to be evaluated, and the amount of credit to be allotted to each. For a restricted-response question calling for three hypotheses, for example, a list of acceptable hypotheses would be prepared, and a given number of scoring points would be assigned to each. For an extended-response question, the major points or aspects of the answer would be outlined. In addition, the relative amount of credit to be allowed for such characteristics as accuracy of the factual information, pertinence of examples, skill of organization, and effectiveness of presentation would be indicated.

Preparing a scoring rubric provides a common basis for evaluating the students' answers and increases the likelihood that our standards for each question will remain stable throughout the scoring. If prepared during the test's construction, such a scoring key also helps us phrase questions that clearly convey the types of answers expected. For a restricted-response essay question, a point might be assigned to each of two or three desired properties of the responses, and a point would be awarded to a student response for each of the desired properties it contained. For an extended-response essay question, a 5-point rating might be used. Five points would be awarded to a response that was well organized and clear and that displayed the type of analysis and reasoning sought by the question. Three points might be awarded for an answer that was clear and adequate but not very compelling. Answers that contained little accurate information and displayed inadequate reasoning might be awarded a single point.

2. Use the scoring rubric that is most appropriate. As discussed previously, two types of scoring rubrics, analytic and holistic, are commonly used with essay questions. Analytic rubrics focus attention on one characteristic at a time and are especially useful in providing students with specific feedback about aspects of their work. Holistic rubrics are likely to be more useful when the focus of the assessment is on overall content understanding rather than writing skill per se.

3. Decide how to handle factors that are irrelevant to the learning outcomes being measured. Several factors influence our evaluations of answers that are not directly pertinent to the purposes of the measurement. Prominent among these are legibility of handwriting, spelling, sentence structure, punctuation, and neatness. We should make an effort to keep such factors from influencing our judgment when evaluating the content of the answers. In some instances, such factors may, of course, be evaluated for their own sake. When this is done, you should obtain a separate score for written expression or for each of the specific factors. As far as possible, however, we should not let such factors contaminate the extent to which our scores reflect the achievement of other learning outcomes.

Another decision concerns the presence of irrelevant and inaccurate factual information in the response. Should you ignore it and score only that which is pertinent and correct? If you do, some students will write everything that occurs to them, knowing that you will sort it out and give them credit for anything correct. This discourages careful thinking and desirable evaluative abilities. On the other hand, if you reduce scores for irrelevant and inaccurate material, the question of how much to lower the score on a given paper is a troublesome one. Probably the best procedure is to decide in advance approximately how much the score on each question is to be lowered when the inclusion of irrelevant material is excessive. The students should then be warned that such a penalty will be imposed.

4. Evaluate all responses to one question before going on to the next one. One factor that contributes to unreliable scoring of essay questions is a shifting of standards from one paper to the next. A paper with average answers may appear to be of much higher quality when it follows a failing paper than when it follows a near-perfect one. One way to minimize this is to score all answers to the first question, reorder the papers to be evaluated, then score all answers to the second question and so on until all the questions have been scored. A more uniform standard can be maintained with this procedure

because it is easier to remember the basis for judging each answer and because answers of various degrees of quality can be more easily compared. When the rating method is used and the responses are placed in several piles on the basis of each answer, shifting standards also can be checked by evaluating each answer a second time and reclassifying it if necessary.

Evaluating all answers to one question at a time helps counteract another type of error that creeps into the scoring of essay questions. When we evaluate all the answers of a single student, the first few answers create a general impression of the student's achievement that colors our judgment of the remaining answers. Thus, if the first answers are of high quality, we tend to overrate the following answers; if they are of low quality, we tend to underrate them. This "halo effect" is less likely when the answers for a given student are not evaluated in continuous sequence.

5. When possible, evaluate the answers without looking at the student's name. The general impression we form about each student during our teaching is also a source of bias in evaluating essay questions. It is not uncommon for a teacher to give a high score to a poorly written answer by rationalizing that "the student is really capable even though she didn't express it clearly." A similar response by a student regarded less favorably will receive a much lower score, with the honest conviction that the student deserved the lower score. This halo effect is one of the most serious deterrents to reliable scoring by classroom teachers and is especially difficult to counteract. See the box "Bluffing: A Special Scoring Problem" for information about a scoring problem unique to essay questions.

When possible, the identity of the students should be concealed until all answers are scored. The simplest way to do this is to have the students put their names on the back of the papers. If a student's identity cannot be concealed because of familiar handwriting, the best we can do is make a conscious effort to eliminate any such bias from our judgment.

6. If especially important decisions are to be based on the results, obtain two or more independent ratings. Sometimes essay questions are included in assessments used to select students for awards, scholarships, special training, and the like. In such cases, two or more competent persons should score the responses independently, and their ratings should be compared. After any large discrepancies have been satisfactorily arbitrated (possibly by a third scorer), the independent ratings may be averaged for more reliable results.

SUMMARY

The essay question is especially useful for measuring those aspects of complex achievement that cannot be measured well by more objective means. These include (a) the ability to supply rather than merely identify interpretations and applications of data and (b) the ability to organize, integrate, and express ideas in a general attack on a problem. Outcomes of the first type are measured by restricted-response questions and outcomes of the second type by extended-response questions.

Although essay questions provide an effective means of measuring significant learning outcomes, they have several limitations: (a) scoring tends to be unreliable, (b) scoring

Bluffing: A Special Scoring Problem

It is possible for students to obtain higher scores on essay question responses than they deserve by means of clever bluffing. This is usually a combination of writing skill, general knowledge, and common "tricks of the trade." Following are some ways that students might attempt to influence the reader and, thus, inflate their grades.

1. Writing something for every question, even if it is only a restatement of the question. (Students figure they might get some credit. Blank spaces get none.)
2. Stressing the importance of the topic covered by the question, especially when short on facts (e.g., "This battle played a significant role in the Civil War.").
3. Agreeing with the teacher's views whenever it seems appropriate (e.g., "The future of mankind depends on how well we conserve our natural resources.").
4. Being a name-dropper (e.g., "This is supported by the well-known experiment by Smith." The reader assumes that the student knows Smith's "well-known" experiment.).
5. Writing on a related topic and fitting it to the question (e.g., prepared to write on President Harry Truman but asked to write about General Douglas MacArthur, the student might start with, "Harry Truman was the president who fired General MacArthur." From then on, there is more about President Truman than General MacArthur.).
6. Writing in general terms that can fit many situations (e.g., in evaluating a short story, the student might say: "This was an interesting story. The characters were fairly well developed, but in some instances more detail would be welcome." This might be called the fortune-teller approach.).

Although bluffing cannot be completely eradicated, carefully phrasing the questions and following clearly defined scoring procedures can reduce it.

is time consuming, and (c) only a limited sampling of achievement is obtained. Because of these shortcomings, essay questions, especially ones requiring extended responses, should be limited to assessing those outcomes that cannot be measured well by objective items.

The construction and scoring of essay questions are interrelated processes that require attention if a valid and reliable measure of achievement is to be obtained. Questions should be phrased so that they measure the attainment of definite learning outcomes and clearly convey to the students the type of response expected. To the extent possible, scoring criteria should be specified in advance. For restricted-response essay questions, scoring rubrics can usually be generated by outlining possible answers deserving full credit and indicating what aspects of the answers are required for different amounts of partial credit. For extended-response essays, a choice between analytic and holistic scoring rubrics should be made. Analytic scoring rubrics have the advantage of providing students with more specific feedback than holistic scoring rubrics. Holistic scoring rubrics can be

developed and applied more rapidly and may correspond closely to grading decisions that need to be made. Available examples of both analytic and holistic scoring rubrics provide useful starting points for developing rubrics for classroom use.

Indicating an approximate time limit for each question and avoiding the use of optional questions also contribute to more valid results. Scoring procedures can be improved by (a) using a scoring rubric, (b) adapting the scoring method to the type of question used, (c) controlling the influence of irrelevant factors, (d) evaluating all answers to each question at one time, (e) evaluating without looking at the students' names, and (f) obtaining two or more independent ratings when important decisions are to be made.

LEARNING EXERCISES

1. In an area in which you are teaching or plan to teach, identify several learning outcomes that can be best measured with essay questions. For each learning outcome, construct two essay questions.
2. Criticize the following essay questions and restate them so that they meet the criteria of a good essay question.
   a. Discuss air transportation.
   b. Do you think the government should spend more on environmental protection?
   c. What is your attitude toward health care reform?
3. For each of the following, would it be more appropriate to use an extended-response question or a restricted-response question?
   a. Compare two periods in history.
   b. Describe the procedure for using a dictionary.
   c. Indicate the advantages of one procedure over another.
   d. Evaluate a short story.
4. Construct an analytic and a holistic scoring rubric for an extended-response essay question that might be used in the grade and content area of most interest to you.
5. What factors should be considered in deciding whether essay questions should be included in a classroom test? Which of the factors are most important?
6. Describe how essay tests might be used to facilitate learning. What types of learning are most likely to be enhanced?

REFERENCES

Applebee, A. N., Langer, J., & Mullis, I. V. S. (1994). NAEP 1992 Writing Report Card. National Center for Education Statistics. Washington, DC. GPO (065-000-00654-5).

Gearhart, M., Herman, J. L., Baker, E. L., & Whittaker, A. K. (1994). *Writing portfolios at the elementary level: A study of methods for writing assessment* (CSE Technical Report 337). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu

Wolf, S. A., & Gearhart, M. (1997). New writing assessments: The challenge of changing teachers' beliefs about students as writers. *Theory Into Practice* 220–230. (Also available as CSE Technical Report 400. Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing. Available at http://www.cse.ucla.edu)

FURTHER READING

Gronlund, N. E., & Waugh, C. K. (2009). *Assessment of student achievement* (9th ed.). Upper Saddle River, NJ: Pearson. Chapter 8, "Writing Supply Items: Short Answer and Essay," discusses the construction and use of essay questions.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development. Presents examples of scoring rubrics and discusses approaches to developing essay assessments.

Regional Educational Laboratories. (1998). *Improving classroom assessment: A toolkit for professional developers.* Available from Regional Educational Laboratories or centrally from Northwest Regional Educational Laboratory, Portland, OR. Includes samples of performance assessments and scoring rubrics.

Welch, C. (2006). Item and prompt development in performance testing. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 303–328), Mahwah, N.J.: Erlbaum.