CHAPTER

9

MEASURING COMPLEX ACHIEVEMENT: THE INTERPRETIVE EXERCISE

Complex achievement includes those learning outcomes based on the higher mental processes, such as understanding, thinking skills, and various problem-solving abilities. Although some aspects of complex achievement require extended constructed responses and other types of performance assessment tasks, other aspects can be measured objectively.

We have already had some experience with measuring complex achievement, as this category encompasses all those learning outcomes requiring more than the mere retention of factual knowledge. The use of the short-answer item to measure problem-solving abilities in mathematics and science, the true–false item to measure the ability to recognize cause-and-effect relationships, and the multiple-choice item to measure various aspects of understanding and application all illustrate the measurement of complex achievement. These illustrations, however, were limited to the use of single, independent test items of the objective type. Greater range and flexibility in measuring complex achievement can be attained not only by moving to the extended response and other performance assessment tasks discussed in Chapters 10 and 11 but also by using more complex forms of objective test items.

A variety of learning outcomes are included in complex achievement. Following are typical examples.

- Ability to apply a principle
- Ability to interpret relationships
- Ability to recognize and state inferences
- Ability to recognize the relevance of information

- Ability to develop and recognize tenable hypotheses
- Ability to formulate and recognize valid conclusions
- Ability to recognize assumptions underlying conclusions
- Ability to recognize the limitations of data
- Ability to recognize and state significant problems
- Ability to design experimental procedures
- Ability to interpret charts, tables, and data
- Ability to evaluate arguments

These and similar learning outcomes have been classified under such categories as understanding, reasoning, critical thinking, scientific thinking, creative thinking, and problem solving. There is general agreement that learning outcomes based on higher-order thinking skills constitute some of the most significant outcomes of education. Given the importance of these complex learning outcomes, it is critical to use a full array of assessment techniques available for measuring those outcomes. The interpretive exercise provides one of those needed techniques. Used wisely, and supplemented by the techniques discussed in Chapters 10 and 11, the interpretive exercise can help ensure that complex learning outcomes are given adequate priority in classroom assessments.

NATURE OF THE INTERPRETIVE EXERCISE

An interpretive exercise (also called "classification exercise," "key-type item," or "master-list item") consists of a series of objective items based on a common set of stimuli. The stimuli may be in the form of written materials, tables, charts, graphs, maps, or pictures. The series of related test items may also take various forms but are most commonly multiple-choice or true–false items. Because all students are presented with a common set of stimuli, it is possible to measure a variety of complex learning outcomes. Students can be asked to identify relationships in data, to recognize valid conclusions, to appraise assumptions and inferences, to detect proper applications of data, and the like.

The common set of materials used in interpretive exercises ensures that all students will be confronted with the same task. It also makes it possible to control the amount of factual information given to them. We can give them as much or as little information as we think desirable in measuring their achievement of a learning outcome. In measuring their ability to interpret mathematical data, for example, we can include the formulas needed or require the students to supply them. In other areas, we can supply definitions of terms, meanings of symbols, and other facts or expect students to supply them. This flexibility makes it possible to measure various degrees of proficiency in any particular area.

FORMS AND USES OF THE INTERPRETIVE EXERCISE

As with other objective items, there are so many forms and uses of the interpretive exercise that it is impossible to illustrate all of them. Here we present examples of this item type as applied to the measurement of complex learning outcomes in a variety of school

subjects at the elementary and secondary levels. Different types of introductory material and different methods of responding also will be used to illustrate the great flexibility of the interpretive exercise. The references at the end of this chapter offer additional illustrative exercises.

Ability to Recognize Inferences

In interpreting written material, it is frequently necessary to draw inferences from the facts given. The following exercise measures the extent to which students are able to recognize warranted and unwarranted inferences drawn from a passage.

EXAMPLE

Directions: Assuming that the information below is true, it is possible to establish other facts using the ones in this paragraph as a basis for reasoning. This is called drawing inferences. There is, of course, a limit to the number of kinds of facts which may be properly inferred from any statement.

By writing the proper symbol in the space provided, indicate that a statement is TRUE if it may be properly inferred from the information given in the paragraph. Indicate that it is UNTRUE if the information given in the paragraph implies that it is false. Indicate that NO INFERENCE can be drawn if the statement cannot be inferred one way or the other. Use only the information given in the paragraph as a basis for your responses. . . .

Use the following symbols in writing your answers:

T—if the statement may be inferred as TRUE.
F—if the statement may be inferred as UNTRUE.
N—if NO INFERENCE can be drawn about it from the paragraph.

PARAGRAPH A

By the close of the thirteenth century there were several famous universities established in Europe, though of course they were very different from modern ones. One of the earliest to be founded was one of the most widely known. This was the University of Bologna, where students from all countries came who wished to have the best training in studying Roman law. Students especially interested in philosophy and theology went to the University of Paris. Those who wished to study medicine went to the Universities of Montpellier or Salerno.

QUESTIONS ON PARAGRAPH A

(T) 1. There were law suits between people occasionally in those days.
(N) 2. The professors were poorly paid.
(F) 3. In the Middle Ages people were not interested in getting education.
(T) 4. There were books in Europe at that time.
(N) 5. Most of the teaching in these medieval universities was very poor.
(N) 6. There was no place where students could go to study.
(F) 7. There were no doctors in Europe at this time.
(F) 8. There was no way to travel during the Middle Ages.
(T) 9. If a student wanted to be a priest, he would probably attend the University of Paris.

(N) 10. There were no universities in Europe before the thirteenth century.
(N) 11. There was only one language in Europe at this time.

Source: From "Selected Items for the Testing of Study Skills" by H. T. Morse and G. H. McCune, 1971, *Bulletin, 15,* 66. Copyright 1971 by National Council for the Social Studies. Used by permission of the publisher.

Ability to Recognize Warranted and Unwarranted Generalizations

The ability to recognize the validity of generalizations is of central importance in the interpretation of data. At minimum, students should be able to determine which conclusions the data support, which the data refute, and which the data neither support nor refute. The data may be in the form of tables, charts, graphs, maps, or pictures, and the test items may be true–false or multiple-choice items. An illustration of recognizing the validity of generalizations is shown in the following example.

EXAMPLE Percentage of population between the ages of 25 and 29 who have completed secondary and college (a bachelor's degree or higher education), by gender in 1980, 1985, 1990, 1995, 2000, and 2005

	Males		Females	
Year	High School	College	High School	College
1980	85.4	24.0	85.5	21.0
1985	85.9	23.1	86.4	21.3
1990	84.4	23.7	87.0	22.8
1995	86.3	24.5	87.4	24.9
2000	86.7	27.9	89.4	30.1
2005	84.9	25.3	87.3	32.0

Source: Data from "The Condition of Education: 2006," Washington, DC: National Center for Education Statistics, U. S. Department of Education, 2006.

Directions: The following statements refer to the data in the table above. Read each statement and mark your answer according to the following key.

Circle:

S if the statement is Supported by the data in the table.
R if the statement is Refuted by the data in the table.
N if the statement is Neither supported nor refuted by the data.

(S) R N 1. The discrepancy in percentage completion of higher education for males and females between the ages of 25 and 29 was smaller in 1995 than it was any of the other years shown in the table.
S R (N) 2. Since 2000, college admissions policies give preferential treatment to female applicants over male applicants.
S R (N) 3. It was more difficult to get into college in the 1980s and 1990s than it is today.
S (R) N 4. When males and females are combined, the percentage for young adults between the ages of 25 and 29 who have completed high school has increased every year over what it was 5 years earlier.

Ability to Recognize Assumptions

Another learning outcome pertinent to the interpretation of various types of information is the ability to identify unstated assumptions that are necessary to a conclusion or course of action. The following item illustrates this type of interpretive exercise.

EXAMPLE Studies have shown that there is a relationship between vocabulary and crime. Crime rates are higher for people with poorly developed vocabularies, and crime rates are lower for people with well-developed vocabularies. Older studies have also shown that there is a positive relationship between the number of years of Latin studied and the size and preciseness of an individual's vocabulary. Conclusion: Crime rates can be lowered by reintroducing the study of Latin in the schools.

Which one of the following assumptions is necessary to reach such a conclusion?

A Correlational methods were used to determine these relationships.
B These reported relationships were statistically significant.
(C) Relationships such as these imply causation.
D Latin scholars have a low crime rate.

Ability to Recognize the Relevance of Information

A learning outcome important to all subject-matter areas and that can be measured at all levels of instruction is the ability to recognize the relevance of information. The exercise presented here was prepared for third-grade students.

EXAMPLE Bill lost his boot on the way to school. He wanted to put a notice on the bulletin board so that the other children could help him find it. Which of the following sentences tell something that would help children find the boot?

Directions: Circle *yes* if it would help. Circle *no* if it would not help.

(yes)	no	1.	The boot was black.
yes	(no)	2.	It was very warm.
(yes)	no	3.	It was for his right foot.
yes	(no)	4.	It was a Christmas present.
yes	(no)	5.	It was nice looking.
(yes)	no	6.	It had a zipper.
(yes)	no	7.	It had a gray lining.

Ability to Apply Principles

The application of principles may be shown in many different ways. In the following example, students are asked to identify principles that explain a situation and to recognize illustrations of a principle.

EXAMPLE Mary Ann wanted her rose bush to grow faster, so she applied twice as much chemical fertilizer as was recommended and watered the bush every evening. About a month later she noticed that the rose bush was dying.

Directions: Which of the following principles is necessary in explaining why the rose bush is dying? If a principle is Necessary, circle **N**; if a principle is Unnecessary, circle **U**.

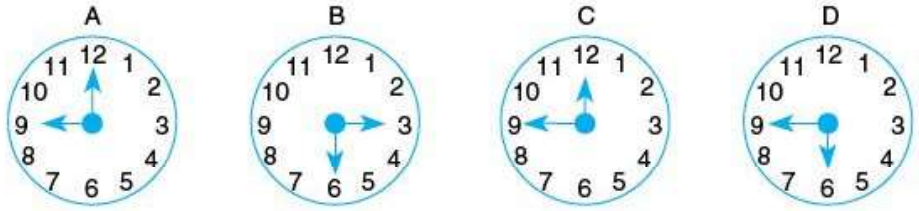
N	(U)	1. A chemical compound is changed into other compounds by taking up the elements of water.
(N)	U	2. Semipermeable membranes permit the passage of fluid.
N	(U)	3. Water condenses when cooled.
(N)	U	4. When two solutions of different concentration are separated by a porous partition, their concentration tends to equalize.

Use of Pictorial Materials

Pictorial materials can serve two useful purposes in interpretive exercises. First, they can help measure a variety of learning outcomes similar to those already discussed simply by replacing the written or tabular data with a pictorial presentation. This use is especially desirable with younger students and when ideas can be more clearly conveyed in pictorial form. Second, pictorial materials can also measure the ability to interpret graphs, cartoons, maps, and other pictorial materials. In many school subjects, these are important learning outcomes in their own right.

The following examples illustrate the use of pictorial materials.

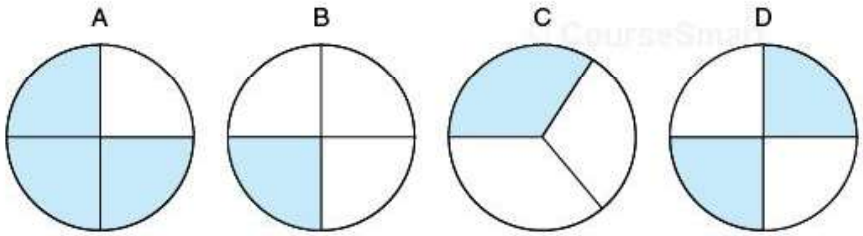
EXAMPLE I



USE ORAL QUESTIONS

What clock shows the time that school starts?	(A) B C D
What clock shows the time closest to lunch time?	A B (C) D
What clock shows half past the hour?	A (B) C D

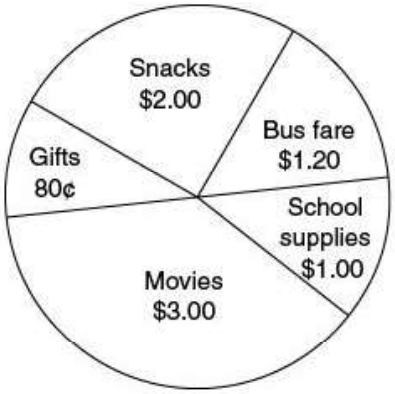
EXAMPLE II



Use Oral Questions

What circle is 1/4 shaded? A (B) C D
What circle is 1/2 shaded? A B C (D)
What circle is *most* shaded? (A) B C D
What circle is *least* shaded? A (B) C D

EXAMPLE III



Above is a graph of Bill's weekly allowance distribution.

1. What is the ratio of the amount Bill spends for school supplies to the amount he spends for movies?

   A 7:2
   (B) 1:3
   C 2:7
   D 3:1

2. What would be the best title for this graph?

   A Bill's weekly allowance
   B Bill's money graph
   (C) Bill's weekly expenditures
   D Bill's money planning

These three examples were designed for use in lower grades. They illustrate the use of pictorial materials that can be drawn by the teacher and items that are useful for measuring rather simple interpretations of concepts and relationships.

Examples IV and V are interpretive exercises designed for higher grade levels. They are included here to illustrate the use of various types of pictorial materials, the measurement of different types of learning outcomes, and the use of both multiple-choice and true–false items. As noted in these examples, the pictures and diagrams used in an interpretive exercise frequently can be obtained from published sources. When this is done, care must be taken in reproducing the pictorial elements to make certain that they are clear and detailed enough for proper interpretation. It is also important, of course, to be aware of

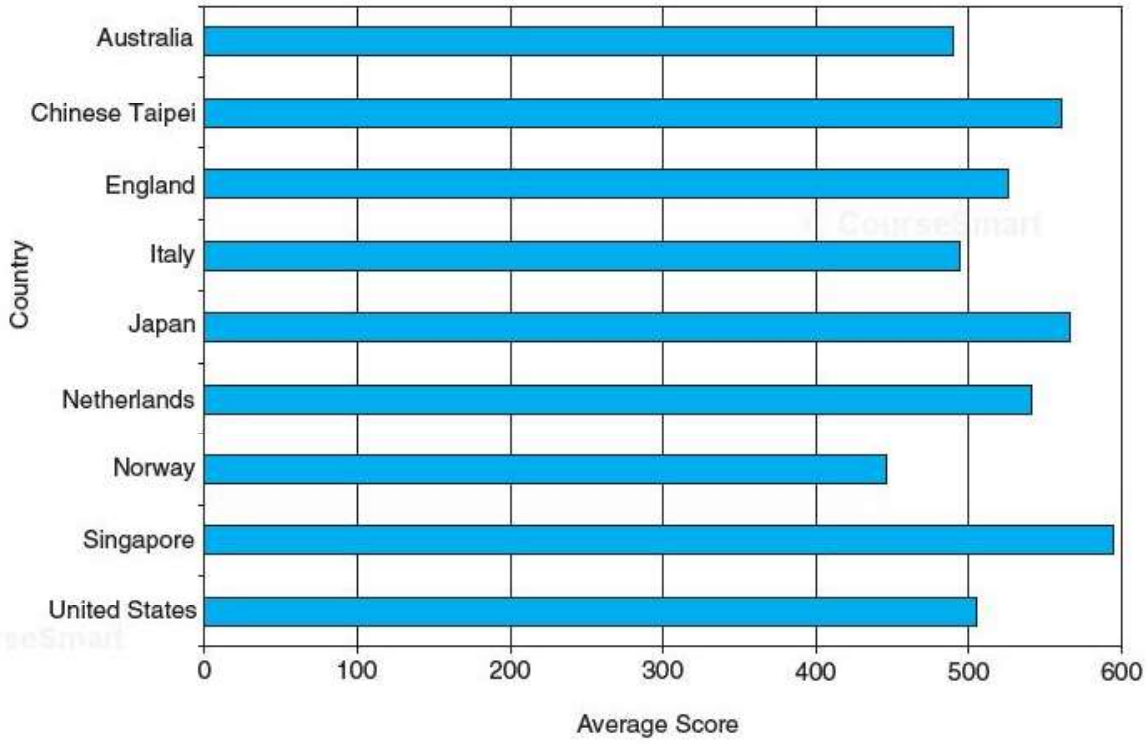
the copyright laws that govern the use of the material. However, there is seldom a problem in obtaining permission to reproduce copyrighted materials for classroom use.

Cartoons like the one in Example IV can be found in newspapers and news magazines. Then simply prepare questions that require the desired interpretations. Either true–false or multiple-choice items might be used with this type of exercise. It is important to select a cartoon that illustrates a concept or principle that is relevant to the learning outcomes to be measured. Interpretive exercises of this type are especially useful in social studies.

EXAMPLE IV

(This item omitted from WebBook edition)

EXAMPLE V TIMSS Average Fourth-Grade Mathmatics, score by Country. (From IEA's TIMSS 2003 International Report on Achievement in the Mathematics Cognitive Domains, Boston College.)



Directions: The following statements refer to the data in the chart above. Read each statement, and mark your answer according to the following key.

Circle:

T—if the data in the chart are sufficient to make the statement True.
F—if the data in the chart are sufficient to make the statement False.
I—if the data in the chart are Insufficient to determine whether the statement is true or false.

T (F) I 1. The average score is lower in the United States than in all but one of the other countries shown.

(T) F I 2. The average score of days is higher in the three Asian countries than in any of the remaining six countries shown.

T F (I) 3. Norwegian students spend fewer hours studying mathematics than students from any of the other countries shown.

A chart like that in Example V is typically easy to prepare but may sometimes be found in published sources. Maps and diagrams that contain data also make effective materials for interpretive exercises.

ADVANTAGES AND LIMITATIONS OF INTERPRETIVE EXERCISES

The interpretive exercise has several advantages. First, the introductory material makes it possible to measure the ability to interpret written materials, charts, graphs, maps, pictures, and other communication media encountered in everyday situations. The rapid expansion of knowledge in every subject-matter area has made it impossible to learn all the important factual information in a given field, which has led to greater dependence on the internet, libraries, reference materials, self-study techniques, and interpretive skills. Second, the interpretive exercise makes it possible to measure more complex learning outcomes than can be measured with the single objective item. Some data are usually necessary if students are to demonstrate thinking and problem-solving skills, and the inclusion of such data is awkward in most kinds of test items. Third, by having a series of related test items based on a common set of data, greater depth and breadth can be obtained in the measurement of achievement skills. Fourth, the interpretive exercise minimizes the influence of irrelevant factual information on the measurement of complex learning outcomes. Students may be unable to demonstrate their understanding of a principle simply because they do not know some of the facts concerning the situation to which they are to be applied. This blocking of response, caused by a lack of detailed factual information not directly pertinent to the purpose of the measurement, can be largely eliminated with the interpretive exercise. In the introductory materials, we can give students the common background of information needed to demonstrate understanding, thinking skills, and problem-solving abilities.

The interpretive exercise is more structured than performance assessment tasks. Whether this is an advantage or disadvantage depends on the specific outcomes to be measured. Students are not free to redefine the problem or to demonstrate thinking skills at which they are most efficient on an interpretive exercise. The series of objective items forces them to use only the mental processes called for. This also makes it possible to measure separate aspects of problem-solving ability and to use objective scoring procedures. This structure has advantages for focusing the task on a specific outcome, such as the ability to identify assumptions underlying conclusions. For other types of outcomes, however, the structure may make it impossible to adequately assess the desired outcome. For example, an interpretive exercise would not be suitable for measuring whether students are able to generate the assumptions underlying a conclusion.

As with all forms of test items, the interpretive exercise has certain limitations. Probably the greatest limiting factor, and one that may have occurred to you as you reviewed the sample items, is the difficulty of construction. Selecting printed materials that are new to the students but relevant to the instructional outcomes requires considerable searching. When pertinent material is found, it usually must be edited and reworked to make it more suitable for testing purposes.

Next, test items must be constructed that demand the specific behaviors indicated in the learning outcomes being measured. The construction process is often circular (i.e., it goes back and forth between revising the introductory material and revising the test items until a satisfactory product is obtained). This entire procedure is time consuming and requires much greater skill than that needed to construct single objective test items. Three positive comments can be made regarding the difficulty of constructing interpretive exercises, however. First, more and more items of this type now appear in various subject-matter fields. The references at the end of this chapter contain numerous examples that may serve as guides to test construction.

Second, the greater instructional emphasis on complex learning outcomes resulting from the use of interpretive exercises offsets the additional effort required in test construction. Finally, the task becomes easier with practice and experience.

A second limitation, especially pertinent when the introductory material is in written form, is the heavy demand on reading skill. The poor reader is handicapped by both the difficulty of the reading material and the length of time it takes to read each test exercise. The first problem can be controlled somewhat by keeping the reading level low and the second by using brief passages. Both of these are only partial solutions, however, because the poor reader will still be at a decided disadvantage. In the primary grades and in classes that contain many poor readers, interpretive exercises might be better limited to the use of pictorial materials.

Compared to the extended essay questions and performance-based assessment tasks discussed in Chapters 10 and 11, the interpretive exercise has two shortcomings as a measure of complex achievement. First, it cannot measure a student's overall approach to problem solving. It is efficient for measuring specific aspects of the problem-solving process, but it does not indicate whether the student can integrate and use these skills when faced with a particular problem. Thus, it provides a diagnostic view of the students' problem-solving abilities in contrast with the holistic view of essay questions or other performance assessment tasks.

Second, because the interpretive exercise normally uses selection items, it is confined to learning outcomes at the recognition level. To measure the ability to define problems, to formulate hypotheses, to organize data, and to draw conclusions, performance assessment tasks must be used. Clearly, the interpretive exercise is not suitable for assessing a student's ability to communicate effectively in writing, perform an experiment, create a work of art, or make an oral presentation to a group. Essay questions and performance assessment tasks are needed for such outcomes. Nonetheless, the interpretive exercise is a valuable technique that can contribute to the valid measurement of complex outcomes.

SUGGESTIONS FOR CONSTRUCTING INTERPRETIVE EXERCISES

The two main tasks in constructing interpretive exercises are (1) selecting appropriate introductory material, and (2) constructing a series of dependent test items. In addition, care must be taken to construct test items that require analyzing the introductory material in terms of complex learning outcomes. The following suggestions will aid in constructing high-quality interpretive exercises.

1. **Select introductory material that is relevant to the objectives of the course.** Interpretive exercises, like other testing procedures, should measure the achievement of specific instructional outcomes. Success in this regard depends to a large extent on the introductory material, as this provides the common basis for the test items. If the introductory material is too simple, then the exercise may become a measure of general information or simple reading skill. On the other hand, if the material is too complex or unrelated to instructional goals, then it may become a measure of general reasoning ability. Both extremes must be avoided. Ideally, the introductory material should be pertinent to the course content and complex enough to evoke the mental reactions specified in the course objectives.

The amount of emphasis given to the various interpretive skills in the course objectives is also important. Care must be taken not to overload the test with interpretive items in any particular area. The selection of introductory material should be guided by the emphasis to be given to the measurement of complex achievement and each type of interpretive skill.

2. **Select introductory material that is appropriate to the students' curricular experience and reading level.** Many complex learning outcomes can be measured with different types of introductory material. The ability to recognize the validity of conclusions, for example, can be measured with written materials, tables, charts, graphs, maps, or pictures. The type used should be familiar to the students so that the nature of the material does not prevent them from demonstrating their achievement of the complex learning outcomes. It would be unfair, for example, to ask students to recognize the validity of conclusions on the basis of data presented in graph form if they had not had experience in interpreting graphs similar to those used in the test. When various types of introductory material will serve a purpose equally well and all are familiar to the students, we favor material that places the least demand on reading skill. For elementary students, pictorial materials are definitely favored. For higher grade levels, pictorial materials and verbal materials with a low vocabulary load and simple sentences are preferred. Although general reading skill is necessary in all written tests, it can become prominent in interpretive exercises unless efforts are made to minimize its influence.

3. **Select introductory material that is new to students.** In order to measure complex learning outcomes, the introductory material must be new. Asking students to interpret materials identical to those used in instruction does not ensure that the exercise will measure anything other than rote memory. Too much novelty, however, must be avoided. Materials that are similar to those used in class but vary slightly in content or form are the most desirable. Such materials can be obtained by modifying selections from textbooks, newspapers, newsmagazines, and various reference materials pertinent to the course content.

4. **Select introductory material that is brief but meaningful.** Another method of minimizing the influence of general reading skill on the measurement of complex learning outcomes is to keep the introductory material as brief as possible. Digests of articles are frequently available and are good raw material for interpretive exercises. If digests are unavailable, the summary of an article or a key passage may be sufficient. In some cases, the relevant information is summarized better in a table, diagram, or picture. In striving for brief introductory material, be careful not to omit elements that are crucial to the interpretive skills being measured. The material also should, of course, be complete enough to be meaningful and interesting to the students.

5. **Revise introductory material for clarity, conciseness, and greater interpretive value.** Although some materials (e.g., graphs) can be used without revision, most selections require adaptation for testing purposes. Technical articles frequently contain long, detailed descriptions of events. On the other hand, news reports and digests of articles are brief but often present exaggerated reports of events to attract the reader's interest. Although such reports provide excellent material for measuring the ability to judge the relevance of arguments, the need for assumptions, the validity of conclusions, and the like, the material must usually be modified to be used effectively.

Revision of the introductory material and construction of the related test items tend to be interdependent procedures. Rewriting material often suggests questions to be used, and the construction of test questions often necessitates revisions of the material. In revising a description of an experiment, for example, assumptions, hypotheses, or conclusions explicitly stated in the description may be deleted and used as a basis for questions. Likewise, a question calling for application of the experimental findings may require the addition of new material to the selection. Thus, the revision of the introductory material and the construction of test items proceed in a circular fashion until a clear, concise interpretive exercise evolves.

6. Construct test items that require analysis and interpretation of the introductory material. There are two common errors in the construction of interpretive exercises that invalidate them as a measure of complex achievement. One is to include questions that are answered directly in the introductory material, that is, asking for factual information explicitly stated in the selection. Such questions measure reading and recall skills. The second error is to include questions that can be answered correctly without reading the introductory material, that is, requiring answers based on general information. These questions measure simple knowledge outcomes.

If the interpretive exercise is to function as intended, it should include only those test items that require students to read the introductory material and to make the desired interpretations. In some instances, the interpretations will require students to supply knowledge beyond that presented in the exercise. In others, the interpretations will be limited to the factual information provided. The emphasis on knowledge and interpretive skill will be determined by the learning outcomes being measured. Regardless of the emphasis, however, the test items should be dependent on the introductory material while at the same time calling forth mental reactions of a higher order than those related to reading comprehension.

7. Make the number of test items roughly proportional to the length of the introductory material. It is inefficient to have students analyze a long, complex selection of material and then answer only one or two questions about it. Although it is impossible to specify the exact number of questions that should accompany a given amount of material, the items presented earlier in this chapter show a desirable balance. Other things being equal, we always favor the interpretive exercise that has brief introductory material and a relatively large number of test items.

8. In constructing test items for an interpretive exercise, observe all pertinent suggestions for constructing objective items. The form of test item used in the interpretive exercise will determine the rules for construction. If multiple-choice or true–false items are used, the suggestions for constructing these item types should be followed. When modified forms are used, suggestions for constructing each of the various types of objective items should be reviewed for their applicability in construction. Freedom from irrelevant clues and technical defects is as important in interpretive exercises as it is in single, independent test items.

9. In constructing key-type test items, make the categories homogeneous and mutually exclusive. The key-type item, which is used frequently in interpretive exercises, is a modified multiple-choice form that uses a common set of alternatives. In this regard, it is also similar to the matching item and so should be constructed in the same way, with special attention devoted to the categories used in the key. All the categories in any one key should be homogeneous; that is, they all should be concerned with similar types of judgment. At the same

time, there should be no overlapping of categories. Each alternative should provide a separate category so that there is a clear-cut system of classification and each item has only one correct answer.

EXAMPLE

The majority of medical researchers agree that exposure to secondhand cigarette smoke is detrimental to health. A number of cities have passed ordinances that prohibit smoking in public buildings. Despite an intensive educational campaign pointing out the dangers of secondhand smoke, many cities do not prohibit smoking in public buildings. Resolved: In the interests of national health, smoking should be prohibited in all public buildings in the United States.

Directions: Read each of the following statements carefully. In front of each statement mark

Key: **A** if the statement supports the resolution.
B if the statement contradicts the resolution.
C if the statement is a fact.
D if the statement is an opinion.

____1. The amount of reduction in exposure to secondhand smoke in cities with ordinances prohibiting smoking in public buildings has not been studied.

(Similar items complete the exercise.)

In this example, the key includes two overlapping categories, one concerned with the relationship of each statement to the resolution and the other with the nature of the statement itself. This makes it impossible to have only one correct answer for each statement. Item 1, for example, would have to be marked category B because it contradicts the resolution and category C because it is a statement of fact.

The key could be improved by limiting the categories to the relevance of the statements to the resolutions, as illustrated in the following key.

EXAMPLE

Key: **A** if the statement supports the resolution.
B if the statement contradicts the resolution.
C if the statement neither supports nor contradicts the resolution.

If judging both the factual nature of a statement and its relevance is important, these two elements can be combined to form discrete categories as follows:

EXAMPLE

Key: **A** if it is a statement of fact that supports the resolution.
B if it is a statement of opinion that supports the resolution.
C if it is a statement of fact that contradicts the resolution.
D if it is a statement of opinion that contradicts the resolution.

The main drawback to combining two types of judgment in one category is the greater complexity of the key. This is especially undesirable with younger students.

In mathematics, a key-type item that has been found to be quite efficient is the quantitative comparison item. The use of a fixed-response format for quantitative comparison items reduces the reading required and makes it possible for students to respond to a larger number of items in a given period of time.

EXAMPLE

(This item omitted from WebBook edition)

10. In constructing key-type test items, develop standard key categories where applicable. Despite the usefulness of the interpretive exercise for measuring complex achievement, classroom teachers have not used it extensively, often because of the difficulty of construction. The popularity of the key-type item in interpretive exercises is probably because it uses a common set of alternatives. This makes it easier to construct than the regular multiple-choice form, which requires a different set of alternatives for each item.

It is often possible to simplify further the construction of key-type interpretive exercises by preparing key categories that can be reused with different content. For example, a learning outcome such as the ability to recognize assumptions might lead to the following key.

EXAMPLE

Key: **A** an assumption that is necessary to make the conclusion valid.
B an assumption that would invalidate the conclusion.
C an assumption that has no bearing on the validity of the conclusion.

This key could be used with a brief description of a situation, a conclusion based on the situation, and a list of assumptions. Both the key and the form of the item could be used repeatedly, with only the content varying. Although selecting new content material is still a problem, the framework of the standard key categories simplifies the process.

Standard key categories, of course, cannot be used in all areas, and their use should not be permitted to determine which learning outcomes receive emphasis. Rather, the time and effort saved by such procedures should free the teacher to explore more creative applications of the interpretive exercise in other areas. See the "Checklist" box for reviewing interpretive exercises.

CHECKLIST

Reviewing Interpretive Exercises

	Yes	No
1. Is this the most appropriate item format to use?	____	____
2. Is the material to be interpreted relevant to the intended learning outcomes?	____	____
3. Is the material to be interpreted appropriate to the students' curricular experience and reading level?	____	____
4. Have pictorial materials been used whenever appropriate?	____	____
5. Does the material to be interpreted contain some novelty (to require interpretation)?	____	____
6. Is the material to be interpreted brief, clear, and meaningful?	____	____
7. Are the test items based directly on the introductory material (cannot be answered without it), and do they call for interpretation (not just recall or simple reading skills)?	____	____
8. Have reasonable numbers of test items been used in each interpretive exercise?	____	____
9. Do the test items meet the relevant criteria of effective item writing?	____	____
10. When key-type items are used, are the categories homogeneous and mutually exclusive?	____	____
11. If revised, are the interpretive exercises still relevant to the intended learning outcomes?	____	____
12. Have the interpretive exercises been set aside for a time before reviewing them?	____	____

SUMMARY

Complex achievement refers to those learning outcomes based on higher mental processes. Such outcomes are classified under various general headings, including understanding, reasoning, thinking, and problem solving. The attainment of goals in these areas can be measured by both objective and subjective means. The most commonly used objective item is the interpretive exercise.

The interpretive exercise consists of a series of objective questions based on written materials, tables, charts, graphs, maps, or pictures. The questions require students to demonstrate the specific interpretive skill being measured. For example, students might be asked to recognize assumptions, inferences, conclusions, relationships, applications, and the like. The structure of the interpretive exercise makes it possible to obtain independent measures of each aspect of thinking and problem solving. Although it is efficient for measuring such learning outcomes, it does not measure a student's ability to integrate and use these skills in a global attack on a problem. Thus, it is limited to a diagnostic analysis of problem-solving skills.

Probably the main reason for not using the interpretive exercise is the difficulty of construction. This process involves (a) selecting appropriate introductory material, (b) revising the material to fit the outcomes to be measured, and (c) constructing a series of dependent test items that call forth the desired behavior. Although these steps are admittedly time consuming, the rewards in improved teaching–learning practices seem to justify the time and effort.

LEARNING EXERCISES

1. What are the advantages of the interpretive exercise over the performance-based assessment for measuring complex achievement? What are the disadvantages?
2. For which types of learning outcomes is the interpretive exercise most likely to be appropriate? Why?
3. Discuss the relative merits of the interpretive exercise and the single-item multiple-choice question. For which situation would each be most useful? What are the limitations of each?
4. Construct one interpretive exercise for each of the following:
   a. Paragraph of written material
   b. Picture or cartoon
   c. Chart or graph
5. What steps would you follow in examining an interpretive exercise to determine whether it had been properly constructed?
6. What are some of the factors to consider when you are deciding whether to use interpretive exercises in a classroom test?

FURTHER READING

Educational Testing Service. (1973). *Multiple-choice questions: A close look.* Princeton, NJ: Author. Illustrates the use of the multiple-choice item for measuring complex achievement in a variety of fields. Maps, graphs, pictures, diagrams, and written materials are used. Each item is followed by a statistical and logical analysis of its effectiveness.

Gronlund, N. E. (2005). *Assessment of student achievement* (8th ed.). Boston: Allyn & Bacon. See Chapter 5, "Writing Selection Items: True–False, Matching, and Interpretive Exercises," for examples of interpretive exercises and their uses.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). New York: Holt, Rinehart & Winston. See Chapter 7, "Writing Objective Test Items: Multiple-Choice and Context-Dependent," for sample interpretive exercises and suggestions for construction.

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 81–129). Washington, DC: American Council on Education. An extended treatment of the topic of item writing. See pages 120–128 for the construction of interpretive exercises.

CHAPTER

10

MEASURING COMPLEX ACHIEVEMENT: ESSAY QUESTIONS

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. 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 it 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 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 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 students' task is simply to demonstrate an understanding of this material and to put it in their own words. That is, the essay question is simply a measure of comprehension. If the essay question presents students with their 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 requires 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 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.

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.

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)

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 model 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, 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: (1) how to construct essay questions that call forth the desired student responses, and (2) 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. When objective items are inadequate for measuring the learning outcomes, however, the use of essay questions can be easily defended despite their limitations. 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 "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.

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 bases 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 defined. 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.

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 . . .

(*Continued*)

(Continued)

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 . . .

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 what type of response is 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

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, then they must be taught the thinking and writing skills needed to express themselves. This means teaching them how to select relevant ideas, compare and relate ideas, organize ideas, apply ideas, infer, analyze, evaluate, and 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, use of supporting evidence, and mechanics of writing 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?	____	____

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.

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 the following seven analytic dimensions.

1. Ideas and Content
2. Organization
3. Voice
4. Word Choice
5. Sentence Fluency
6. Conventions
7. Citing Sources

Scoring rubrics for 6-point ratings are available on-line at the Oregon Department of Education Web site at http://www.ode.state.or.us/teachlearn/testing/scoring/guides/2006-07/asmtwriscorguide0607eng.pdf. The analytic scoring rubrics are presented for the seven dimensions or "traits." The specification of a score of 6 on the Organization dimension is

shown in the box showing a sample scoring rubric. Similar descriptions are given for score points of 1, 2, 3, 4, and 5 for this and the other six 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 Oregon Department of Education 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 specific details • Details usually clear	• Moderate elaboration of main point	• Some minor errors • One or two major errors

(*Continued*)

Table 10.2 (*Continued*)
Example analytic scales for expository essays or descriptive summaries

Score	*General Impression*	*Focus/ Organization*	*Language*	*Elaboration*	*Mechanics*
		• Some orientation of reader	• Generally clear images • Details usually clear		• 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 digression • 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 main point or central statement	• Many major errors causing reader confusion

Source: Adapted from Gearhart et al. (1994).

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.

Example of the Oregon Department of Education Scoring Rubric to Be Considered a "6" on the Organization Dimension

"The organization enhances the central idea(s) and its development. The order and structure are compelling and move the reader through the text easily. The writing is characterized by

1. Effective, perhaps creative, sequencing and paragraph breaks: the organizational structure fits the topic, and the writing is easy to follow.
2. A strong, inviting beginning that draws the reader in and a strong, satisfying sense of resolution or closure.
3. Smooth, effective transitions among all elements (sentences, paragraphs, ideas).
4. Details that fit where placed."

Narrative essay rubrics for five analytic dimensions were developed by Gearhart, Herman, Baker, and Whittaker (1994) and used by primary classroom teachers in several studies (see Wolf & Gearhart, 1997). The five dimensions are as follows:

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; or see the home page for the Center for Research on Evaluation, Standards, and Student Testing [CRESST] at http://www.cse.ucla.edu).

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

A sample CRESST scoring rubric for use in making holistic ratings of the quality of explanations is shown in Table 10.4.

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, Langer, and Mullis (1994, p. 204).

Table 10.4
Example of CRESST scoring rubric for holistic rating of overall quality of an explanation, grade 10

Score	*Description*
5	The student is extremely knowledgeable about the topic.
This is the highest rating	The student demonstrates in-depth understanding of the relevant and important ideas. The student includes the important ideas related to topic and shows a depth of understanding of important relationships. The answer is fully developed and includes specific facts or examples. The answer is organized somewhat around big ideas, major concepts/principles in the field. The response is exemplary, detailed, and clear.
4	The student is knowledgeable about the topic. The student has a good understanding of the topic. The student includes some of the important ideas related to the topic. The student shows a good understanding of the important relationships. The answer demonstrates good development of ideas and includes adequate supporting facts or examples. The answer may demonstrate some organization around big ideas, major concepts/principles in the field. The response is good, has some detail, and is clear.

(*Continued*)

Table 10.4 (*Continued*)
Example of CRESST scoring rubric for holistic rating of overall quality of an explanation, grade 10

Score	*Description*
3	The student demonstrates some knowledge and understanding of the topic.
This is the middle score of the scale.	The overall answer is OK but may show apparent gaps in his/her understanding and knowledge.
	The student includes some of the important ideas related to the topic.
	The student shows some (but limited) understanding of the relationships.
	The answer demonstrates satisfactory development of ideas and includes some supporting facts or examples.
	The response is satisfactory, containing some detail, but the answer may be vague or not well developed and may include misconceptions or some inaccurate information.
2	The student has little knowledge or understanding of the topic.
	The student may include an important idea, part of an idea, or a few facts but does not develop the ideas or deal with the relationships among the ideas.
	The response contains misconceptions, inaccurate, or irrelevant information.
	The student may rely heavily on the group activity.
	The response is poor and lacks clarity.
1	The student shows no knowledge or understanding of the topic.
	The student either:
	(1) writes about the topic using irrelevant or inaccurate information
	(2) recalls the steps of the Group Activity in Part II of the performance assessment, adding no new or relevant information and showing no understanding of how the activity relates to the general topic.
0	The student either:
	(1) left the answer blank
	(2) wrote about a different topic
	(3) wrote "I don't know."

Source: CRESST: http://www.cse.ucla.edu.

SUGGESTIONS FOR SCORING ESSAY QUESTIONS

Improving the reliability of scoring answers to 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 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)

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.

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 certain limitations: (a) Scoring tends to be unreliable, (b) scoring 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 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*. Washington, DC: National Center for Education Statistics, 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. Available: 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, 36,* 220–230. (Also available as CSE Technical Report 400). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing. Available: http://www.cse.ucla.edu.

FURTHER READING

Gronlund, N. E. (2005). *Assessment of student achievement* (8th ed.). Boston: Allyn & Bacon. Chapter 6, "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.* Portland, OR: Regional Educational Laboratories, or available centrally from Northwest Regional Educational Laboratory. 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* (Chapter 13, pp. 303–327). Mahway, NJ: Lawrence Erlbaum. Provides guidance for improving the quality of essay prompts and gives examples of scoring rubrics.

CHAPTER

11

MEASURING COMPLEX ACHIEVEMENT: PERFORMANCE-BASED ASSESSMENTS

Essay tests are the most common example of a performance-based assessment, but there are many others, including artistic productions, experiments in science, oral presentations, and the use of mathematics to solve real-world problems. The emphasis is on doing, not merely knowing—on process as well as product.

Essay tests are an example of one type of performance assessment, but there are many aspects of writing that are not tapped within the constraints of the normal essay test. Choosing a topic, identifying an audience, gathering information, preparing drafts, seeking critiques, and revising are all important aspects of writing that are not measured by the usual essay test. Moreover, writing is not the only type of performance outcome we need to assess. Many highly valued learning outcomes emphasize the actual performance of tasks in realistic settings. This is obvious in the case of art or music and for vocational or industrial education courses, such as auto repair, woodworking, or word processing. It is also true for mathematics, science, social studies, and foreign languages. In each case, performance-based assessments are needed to measure some of the desired learning outcomes.

For example, although knowledge of vocabulary and grammar in a foreign language can be measured with the various forms of paper-and-pencil tests, speaking skills cannot. Oral performance is required to assess a student's spoken communication skills in a foreign language. Similarly, the assessment of a student's ability to make observations, formulate hypotheses, collect data, and draw valid scientific conclusions may require the use of performance assessments. The use of mathematics to solve meaningful real-world problems and to communicate solutions to others may also be best assessed by the use of performance tasks in realistic settings.

Performance assessments provide a basis for teachers to evaluate both the effectiveness of the ***process*** or procedure used (e.g., approach to data collection or manipulation of instruments) and the ***product*** resulting from performance of a task (e.g., completed report of results or completed artwork). Unlike simple tests of factual knowledge, there is unlikely to be a single right or best answer. Rather, there may be multiple performances and problem solutions that would be judged to be excellent. Problem formulation, the organization of ideas, the integration of multiple types of evidence, and originality are all important aspects of performance that may not be adequately assessed by paper-and-pencil tests.

TYPES OF PERFORMANCE-BASED ASSESSMENT

Performance assessments are also sometimes referred to as "authentic assessments" or "alternative assessments," but the terms are not interchangeable. "Alternative assessment" highlights the contrast to traditional paper-and-pencil tests, whereas "authentic assessment" emphasizes the practical application of the tasks in real-world settings. We prefer the label "performance assessment" because it is more descriptive than "alternative assessment" and less pretentious than "authentic assessment."

Authenticity is a matter of degree. A highly authentic assessment of communication skills in German, for example, might involve listening to the verbal interactions of a student when visiting Germany; but such an assessment obviously would lack practicality for the teacher of a typical German class. Simulated spoken interactions between the teacher and a student or among students, although not quite as authentic, are much more practical. In either case, the focus of the assessment is on the student's performance in communicating in German.

Although authenticity is usually only approximated, it is an important goal of performance assessment. Providing realistic contexts can make problems more engaging for students and help the teacher evaluate whether a student who can solve a problem in one context can solve it in another. Hence, it is desirable to increase the authenticity of tasks to whatever extent possible.

Like essay questions, performance assessments should be used primarily to measure those learning outcomes that cannot be measured well by objective test items. Objective test items are generally more efficient and more reliable for measuring factual knowledge and the ability to solve well-structured problems (e.g., solve a quadratic equation). Performance assessments are better suited for applications with less-structured problems where problem identification; collection, organization, integration, and evaluation of information; and originality are emphasized (e.g., where is the best place to locate a restaurant?). They are also essential for learning outcomes that involve the creation of a product (e.g., a typed letter or a painting) or an oral or physical performance (e.g., the presentation of a speech, the repair of an engine, or the use of a scientific instrument).

Hands-on performance tasks that require students to manipulate objects, measure outcomes, and observe results of experimental manipulations are sometimes essential to capture the full array of skills needed to perform "authentic" tasks. This is obvious in the case of a driving test or a performance test for a dentist, but it may also be true in science and other areas. Research has shown that computer simulations of tasks in science

sometimes may be good substitutes for actual hands-on performance of the task, but in other instances even high-fidelity simulations may have relatively poor relationships for hands-on performance. Poor relationships between simulations and actual hands-on performance occur most commonly when the manipulation of apparatus (e.g., mixing a compound or taking a measurement) is an integral part of the task.

Performance tasks can vary substantially in the degree to which performance is restricted. A word-processing test, for example, might be completely constrained with regard to format and content of a letter to be typed. The task of creating a sculpture might be almost completely unconstrained with regard to the approach a student might take or the nature of the product produced. Most performance tasks fall in between these extremes.

Restricted-Response Performance Tasks

A restricted-response performance task is usually relatively narrow in definition. The instructions are generally more focused than extended-response performance tasks, and the limitations on the types of performance expected are likely to be indicated.

Restricted-response performance tasks sometimes start with a simple multiple-choice or short-answer question, such as the one in Figure 11.1. Those questions are then extended by asking for an explanation of the answer and sometimes an explanation for why the other answers were not selected. Often, different answers in the first part of the task could be given full credit if the explanation provided sound reasoning to defend the choice.

EXAMPLES

Type a letter of application for a job.

Read aloud a section of a story.

Use various combinations of five straight pieces of plastic to construct as many different triangles as you can and record the perimeters of each.

Determine which of two liquids contains sugar and explain what results support your conclusion.

Construct graphs of the average amount of rainfall per month for two cities.

Request aloud directions to the train station in French.

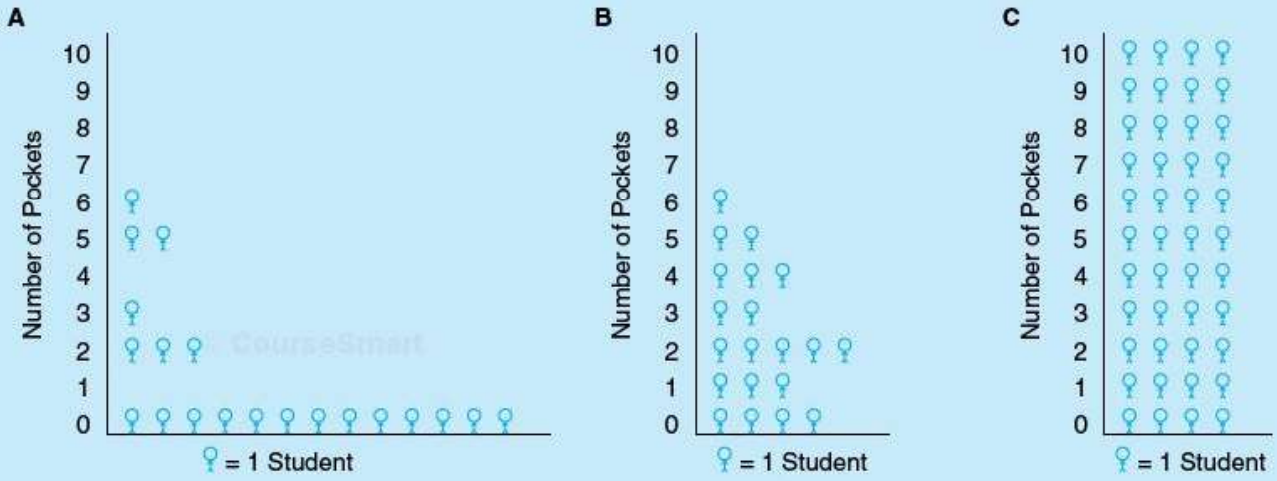
Write the names of the countries in the appropriate areas of a blank map of Europe.

Sara knows that half the students in her class were invited to Kim's birthday party. Also, half were invited to Julie's party. Sara thinks that these figures add up to 100%, so she thinks she will surely be invited to one of the parties. Explain why Sara is wrong. If possible use a diagram in your explanation.*

*Adapted from a task used in the California Assessment program.

If the explanation parts of the task in Figure 11.1 were omitted, there would be no way to determine the basis for a student's choice of one of the three figures. Even if students selected the preferred choice (B), you would not know whether they did so for a sound reason or whether they simply guessed. Nor would you know whether they were attentive to the fact that graph C is impossible because it depicts 44 students when there were only 20 students in Mr. Pang's class.

There are 20 students in Mr. Pang's class. On Tuesday most of the students in the class said they had pockets in the clothes they were wearing.



Which of the graphs most likely shows the number of pockets that each child had?_____

Explain why you chose that graph.

Explain why you didn't choose the other two graphs.

Figure 11.1
Example of stimulus material for a mathematics problem administered at grade 4 in the 1992 National Assessment of Educational progress

Source: NAEP 1992: Mathematics Report Card for the Nation and the States (p. 49) by I. V. S. Mullis, J. A. Dossey, E. H. Owen, and G. W. Phillips, 1993, Washington, DC: U.S. Department of Education. Report No. 23-ST02.

As is true of many tasks that are called performance assessment tasks, the example in Figure 11.1 is, of course, a type of essay question. No real manipulation or hands-on activity is involved. A task such as the one in the figure might readily be adapted to a classroom assessment activity that involved data collection and graphing. Children, for example, might each be asked to count the number of pockets in the clothes they were wearing. Those numbers could be reported, and each student could construct a graph. Separate graphs for boys and for girls in the classroom might also be constructed. They might then be asked to write a description of the graph they constructed before being presented with a task like that in Figure 11.1.

A variety of tasks may be used to assess the skills young students have at making and recording observations, summarizing the observations, and reaching conclusions. In one such task, students were instructed in how to find and count their pulse. They were then asked to count the number of pulses in each of four segments of 15 seconds where the teacher looked at a stopwatch and gave instructions to start and stop. After recording the four initial segments, students were told to jump up and down for 1 minute. After exercise, children were asked to count and record their pulse for four additional 15-second periods. Next, a second period of jumping for 2 minutes was required, followed by four more recordings of 15-second segments. Students were asked to construct a table and a graph reporting the results and describe what the results showed when the initial four recordings were compared to those following the first and second rounds of exercise. Finally, they were asked to explain what they observed.

The relative advantages and disadvantages of restricted performance tasks parallel those of restricted essay questions. They are generally more structured and require less time to administer than extended-response performance tasks. The shorter administration time makes it possible to administer more tasks and thereby gain broader coverage of the content domain. The greater degree of structure makes the task easier to score. On the other hand, the structure makes the tasks less valuable for measuring student skills, such as approaches to ill-structured problems, integration of information, and originality. Extended performance tasks are better suited for such outcomes.

Extended Performance Tasks

The extended performance task may require students to seek information from a variety of sources beyond those provided by the task itself. For example, students may need to use the library, make observations, collect and analyze data in an experiment, conduct a survey, or use a computer or other types of equipment. They may have to identify which aspects of the task are most relevant. The process or procedures that they use may be observed and be an important part of the assessment. The product that is produced may take a variety of forms, such as the construction and presentation of graphs or tables, the use of photographs or drawings, or the construction of physical models. Products may be developed over the course of several days and include opportunities for revision or modification. This freedom enables students to demonstrate their ability to select, organize, integrate, and evaluate information and ideas. The price of these gains includes the loss of efficiency, possible loss of breadth of coverage of the content domain, and greater difficulty in rating performance.

EXAMPLES

Prepare and deliver a speech to persuade people to take actions to protect the environment.

Hog is a game played with dice. The goal is to get the largest possible score. You may roll any number of dice out of a large cup. If none of the numbers is a 1, then the score for the roll is the sum of the numbers rolled. If a 1 is obtained on any of the dice, the score for the roll is zero. What number of dice do you think it best to roll? Defend your decision (Mathematical Sciences Education Board, 1993).

Write a computer program in BASIC that will sort a list of words alphabetically.

Design and carry out an investigation to estimate the acceleration, *a,* of a falling object such as a baseball. Describe the procedure used, present the data collected and analyzed, and state your conclusions.

Read an abridged version of the Lincoln–Douglas debates. Imagine that you were living then and heard the debates. Write a letter to a friend explaining the historical issues addressed and their importance in terms of what you know about the problems facing the nation at the time of the debates (Baker, Aschbacher, Niemi, & Sato, 1992).

Performance assessments require students to demonstrate skills by actually performing. They involve doing rather than just knowing about, and there are sometimes important differences between the two. For example, a guitar player may know which frets to press the strings against for a particular chord without being able to perform the task smoothly to produce the desired sound. Similarly, a computer programmer may know the function of various needed commands without being able to produce a correctly working program to perform a specific task, or a science student may know the parts and functions of an instrument without being able to use it properly to obtain the information needed to solve a problem. Performance assessments are needed to observe and evaluate such skills. They also communicate the message that actual performance is important.

A performance assessment task used in the 1996 NAEP Science Assessment at grade 4 is shown in Figure 11.2. As can be seen, this task requires students to do simple manipulations, to measure and record the outcomes of placing the pencil and thumbtack in the different bottles of water, to draw conclusions about the "mystery water," and to make predictions about the effects of adding salt to a solution. In this example, the manipulations, observations, and measurements are relatively simple, but these basic skills are critical in many settings and are not well assessed in a purely paper-and-pencil assessment.

The effective use of performance assessments requires careful attention to task selection and to the ways performances will be scored. Care needs to be taken in the identification of the complex skills we want to measure, in the construction of tasks that will require students to demonstrate those skills, and in the evaluation of the resulting process and/or product. Without careful attention to these aspects of the assessment, it is unlikely that the effort will yield adequately reliable or valid measures of the complex skills that are being sought.

As the name suggests, performance assessments measure the ability of students to perform tasks that correspond to important instructional objectives. Restricted performance tasks generally focus on specific skills (e.g., reading a passage aloud). Extended performance tasks are more likely to involve problem solving and the integration of a variety of skills and understandings. A comparison of the types of complex learning outcomes measured by each of these types of performance tasks is presented in Table 11.1.

FLOATING PENCIL

Using a Pencil to Test Fresh and Salt Water

You have been given a bag with some things in it that you will work with during the next 20 minutes. Take all of the things out of the bag and put them on your desk. Now look at the picture below. Do you have everything that is shown in the picture? If you are missing anything, raise your hand and you will be given the things you need.

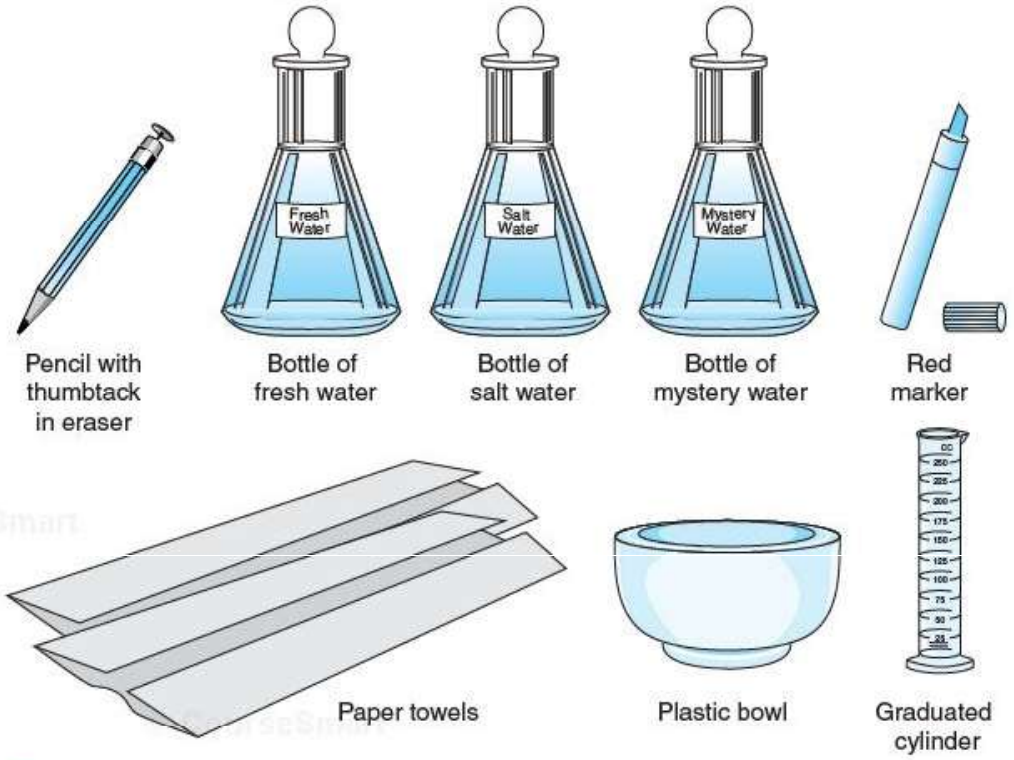


Figure 11.2
Example of hands-on science performance assessment task used at grade 4 in the 1996 National Assessment of Educational Progress

Source: From *NAEP 1996 Science: Report Card for the Nation and the States* by C. Y. O' Sullivan, C. M. Reese, and J. Mazzeo, 1997, Washington, DC: U.S. Department of Education.

ADVANTAGES AND LIMITATIONS OF PERFORMANCE ASSESSMENTS

Advantages

A major advantage of performance assessments is that they can clearly communicate instructional goals that involve complex performances in natural settings in and outside of school. By using tasks that require performances that correspond as closely as is feasible to major instructional objectives, they provide instructional targets and

Table 11.1
Types of performance tasks

Type of Task	*Examples of Complex Learning Outcomes That Can Be Measured*
Restricted-response performance task	Ability to • read aloud • ask directions in a foreign language • construct a graph • use a scientific instrument • type a letter
Extended-response performance task	Ability to • build a model • collect, analyze, and evaluate data • organize ideas, create visuals, and make an integrated oral presentation • create a painting or perform with a musical instrument • repair an engine • write a creative short story

thereby can encourage the development of complex understandings and skills. Often, performance assessment tasks are indistinguishable from good instructional activities.

A second advantage of performance assessments is that they can measure complex learning outcomes that cannot be measured by other means. As has already been stated, knowing how to do something is not the same as being able to do it, much less do it well. Thus, a paper-and-pencil test that measures what a student knows about effective public speaking, for example, does not provide a measure of the student's ability to deliver an effective speech.

A third advantage of performance assessments is that they provide a means of assessing **process** or procedure as well as the **product** that results from performing a task. For example, by observing students while they are conducting a laboratory experiment, strengths and weaknesses in the use of equipment and in technique can be assessed, as can success in completing the experiment and the strength of reasoning provided to support conclusions.

A fourth advantage of performance assessments is that they implement approaches that are suggested by modern learning theory. Rather than viewing students as recipients of discrete bits of knowledge, modern learning theory conceives of students as active participants in the construction of meaning. According to this view, new information must be actively transformed and integrated with a student's prior knowledge. High-quality performance-based assessments take student background knowledge into account and engage students in the active construction of meaning.

Limitations

The most commonly cited limitations of performance assessments parallel those cited for essay questions. Unreliability of ratings of performances across teachers or across time for the same teacher is clearly a limitation. Careful attention to the learning outcomes that the task is intended to assess and to the scoring rubrics that will be used in rating the performances is required both at the time tasks are developed and at the time performances are rated to minimize this limitation. Although the judgmental scoring of complex performances will always include some uncontrollable variations, the scoring reliability, the comparability of scores assigned to the performances of different students, and hence the fairness of the assessment can be greatly increased by clearly defining the outcomes to be measured, properly framing the tasks, and carefully defining and following rubrics for scoring performances.

Another limitation of extended performance assessments is their time-consuming nature. Because a substantial amount of time may be required to allow students to have an adequate opportunity to perform each task, relatively few extended performance assessments can be obtained within a reasonable amount of time. There is considerable evidence that performance on one task provides only a relatively weak basis for generalizing to performances on other tasks intended to assess common or related learning outcomes. Thus, solid generalization to a larger domain of outcomes requires the use of multiple tasks. Overcoming the limitation of weak generalization of performance across tasks requires the accumulation of information from performances on different tasks during the course of the year. Justification for the devotion of the required amount of instructional time to the assessments requires that the tasks provide students with good learning opportunities as well as assessment results.

SUGGESTIONS FOR CONSTRUCTING PERFORMANCE TASKS

The development of high-quality performance assessments that effectively measure complex learning outcomes requires attention to task development and to the ways in which performances are scored. We begin with a consideration of ways to improve the development of tasks and then suggest ways to improve scoring.

1. **Focus on learning outcomes that require complex cognitive skills and student performances.** It is important that tasks be interesting, but that is not sufficient. Tasks need to be developed or selected in light of important learning outcomes. Because performance-based tasks generally require a substantial investment of student time, they should be used primarily to assess learning outcomes that are not adequately measured by less time-consuming approaches.

2. **Select or develop tasks that represent both the content and the skills that are central to important learning outcomes.** Current conceptions of learning stress the interdependence of content and skills. Problem solving in one subject-matter area is not the same as it is in another area. Debating a political issue in social studies is different than debating the effectiveness of a piece of literature. In each case, the content and process are interdependent.

Thus, it is important to specify the range of content and resources students can use in performing a task. Past class assignments provide one natural basis for specifying content, but for many tasks it will be desirable to allow students the opportunity to do additional research to expand their knowledge base. In any event, the specification of assumed content understandings is critical to ensuring that a task functions as intended.

3. **Minimize the dependence of task performance on skills that are irrelevant to the intended purpose of the assessment task.** The key here is to focus on the intention of the assessment. Although both the ability to read complicated texts and the ability to communicate clearly are important learning outcomes, they are not necessarily the intent of a particular assessment. Reading ability, for example, might be irrelevant for an assessment that is intended to measure a student's ability to use mathematics to solve a practical problem (e.g., determine how much and what type of lumber to buy to build a clubhouse with specified features). However, if the task is presented in a way that requires substantial reading, then this factor may add to task difficulty for some students but not for others and thereby reduce the validity of the intended interpretation of the results. This irrelevant source of difficulty would also undermine the fairness of the assessment especially for students with learning disabilities or who are learning English as a second language. On the other hand, writing skills might be an intended part of a mathematics task where a goal of the assessment was to measure a student's ability to communicate mathematical reasoning and results.

4. **Provide the necessary scaffolding for students to be able to understand the task and what is expected.** Challenging tasks often involve ambiguities and require students to experiment, gather information, formulate hypotheses, and evaluate their own progress in solving a problem. However, problems cannot be solved in a vacuum. Students need to have the prior knowledge and skills required to address the problem. These prerequisites can be a natural outcome of prior instruction or may be built in to the task. Preassessment activities, for example, can be used not only to introduce a task but also to ensure that students have the prior knowledge essential for the task and are familiar with the materials or equipment that they need to use. It is important to ask: What prior knowledge and skills are assumed in order to perform the task?

5. **Construct task directions so that the student's task is clearly indicated.** Vague directions can lead to such a diverse array of performances that it becomes impossible to rate them in a fair or reliable fashion. By design, many performance-based tasks give students a substantial degree of freedom to explore, approach problems in different ways, and develop novel solutions. Such intended task characteristics, however, are not an excuse for vague directions. In the task shown in Figure 11.3, students need to experiment and decide on the placement of objects into categories on their own. They also have to construct an explanation for the classification they provide, but the task of using the magnet to test the items, the classification of objects into two categories, and the need to explain the difference between the objects in the two categories are made explicit.

6. **Clearly communicate performance expectations in terms of the scoring rubrics by which the performances will be judged.** Specifying the criteria to be used in rating performance helps clarify task expectations for a student. Explaining the criteria that will be used in rating performances not only provides students with guidance on how to focus their efforts but also helps convey priorities for learning outcomes.

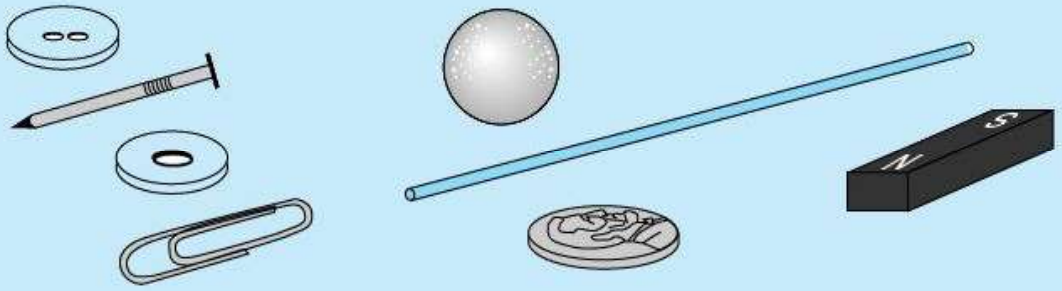
Magnet

Task Descriptor

To use a magnet to identify magnetic and nonmagnetic items and then to explain the difference between them.

Equipment/Material

A magnet and the following seven objects: plastic button, iron or steel washer, steel paper clip, iron nail, glass marble, plastic rod, and copper coin.



Student Instructions

Test the objects with the magnet and divide them into two groups. List the objects in the two groups and explain what makes the objects in the two groups different.

Scoring Scheme

Credit was given for grouping the objects correctly. Four categories of explanations were recorded: namely, that one group was made of iron or steel, that one group was attracted by the magnet, that one group was made of iron and steel and was attracted by the magnet, and any other explanation.

Figure 11.3
Example of performance assessment task in science

Source: Performance Assessment: An International Experiment by B.M. Sample, 1992, Princeton, NJ: Educational Testing Service, Report No. 22-Caep-06. Copyright 1992 by Educational Testing Service. Reprinted by permission.

Listing attributes such as appropriate symbol use, accuracy of information and scale, and ease with which the map can be read makes the rating criteria explicit. It also highlights the learning outcomes that are considered important for the task in the following example.

EXAMPLE Construct a weather map. Your map will be evaluated for accuracy of information and scale, for appropriate use of symbols, and for the ease with which it can be read.

PERFORMANCE CRITERIA

Richard Stiggins (1987) has persuasively argued that the specification of performance criteria is the most important aspect of developing effective performance assessments. He suggests imagining the feedback that would be provided to a student who performed poorly before the task is administered. His rationale for focusing on the criteria to be used is straightforward: "If you do not have a clear sense of the key dimensions of sound performance—a vision of poor and outstanding performance—you can neither teach students to perform nor evaluate their performance."

The criteria to be used in judging student performance are critical for reliable, fair, and valid assessment, and the specification of the criteria should begin at the time the tasks are being selected or developed. Both the teacher and the student need to understand the criteria that will be used to judge performance. As was just noted, criteria help clarify the task expectations for students, and they communicate learning goals and standards. In addition, they guide the judgment process in ways that enhance reliability, fair treatment of each performance, and the validity of conclusions about each student's achievement.

The two main ways of guiding judgments of both the process used in performing a task and any product resulting from that performance are **scoring rubrics/rating scales** and **checklists**. We begin with scoring rubrics and rating scales and then turn to a consideration of checklists.

SCORING RUBRICS AND RATING SCALES

As was discussed in Chapter 10, a scoring rubric is a set of guidelines for the application of performance criteria to the responses and performance of students. A scoring rubric typically consists of verbal descriptions of performance or aspects of student responses that distinguish between advanced, proficient, partially proficient, and beginning levels of performance. Both analytic (Table 10.2) and holistic (Tables 10.3 and 10.4) scoring rubrics were illustrated in Chapter 10.

The analytic scoring rubric requires the identification of different dimensions or characteristics of performance that are rated separately. For example, a mathematics task might be rated in terms of the accuracy of the calculations and the clarity of the explanation. A written report on the results of a science experiment might be rated on factual accuracy, quality of analysis, and the degree to which conclusions were justified. A literary criticism might be rated for organization, quality of ideas, clarity of expression, and mechanics. An oral presentation might be rated both for the substantive quality of the report and for the effectiveness of the presentation.

A holistic rubric provides descriptions of different levels of overall performance. Holistic rubrics are efficient and correspond more directly to global judgments required in the assignment of grades, but they do not provide students with specific feedback about the strengths and weaknesses of their performance as is provided by analytic rubrics.

Rating scales are often limited to making quality judgments (e.g., excellent, good, fair, or poor) or scaled frequency judgments (e.g., always, frequently, sometimes, or never) for each level. As is illustrated in some of the following examples, however, the distinction between scoring rubrics and rating scales is often blurred by adding the descriptions of a rubric to the judgmental qualities of a rating scale.

As is illustrated in Figure 11.4, a scoring rubric may include a rating scale (excellent, good, and so on) but may also provide descriptions of characteristics or performance corresponding to each point on the scale. A scoring rubric makes explicit the criteria that are used to rate performance. Generic scoring rubrics are available that can be readily adapted for use in rating performance on a variety of tasks. Generic scoring rubrics, such as the one shown in Figure 11.4, provide a useful starting place for many assessments. The distinctions between the levels can be made more specific by considering the specific task and likely features that would distinguish between exemplary performance and competent performance or between satisfactory performance with minor flaws and performance that has serious flaws. For example, lists of minor and major flaws might be constructed for a specific task. In a similar fashion, common misconceptions that are anticipated in response to a particular task might be listed.

The number of levels and the verbal descriptions used to guide the scoring may vary from situation to situation. For the hands-on science task involving the floating pencil

Quality of Explanation

6 = Excellent explanation (complete, clear, unambiguous)
5 = Good explanation (reasonably clear and complete)
4 = Acceptable explanation (problem completed but may contain minor flaws in explanation)
3 = Needs improvement (on the right track but may contain serious flaws; demonstrates only partial understanding)
2 = Incorrect or inadequate explanation (shows lack of understanding of problem)
1 = Incorrect without attempt at explanation

Separate Ratings of Answer and Explanation

Answer
4 = Correct
3 = Almost correct or partially correct
2 = Incorrect but reasonable attempt
1 = Incorrect with no relationship to the problem
0 = No answer

Explanation
4 = Complete, clear, logical
3 = Essentially correct but incomplete or not entirely clear
2 = Vague or unclear but with redeeming features
1 = Irrelevant, incorrect, or no explanation

Figure 11.4
Examples of generalized scoring rubrics for mathematics problems

shown in Figure 11.2, for example, the separate scoring rubrics were used for each part of the response. For the part of the task where the student was supposed to identify the mystery water and explain how they could "tell what the mystery water is," student responses were scored using a rubric with three levels:

Complete: Student stated that "the mystery water was fresh water and gave a satisfactory explanation that referred to observations made doing the hands-on task" (O'Sullivan, Reese, & Mazzeo, 1997, p. 44).

Partial: Student stated that the water was fresh but did not support the choice with direct reference to observations from the hands-on task.

Incorrect: Student gave the wrong answer or gave contradictory explanation for the choice of the correct answer of fresh water.

EXAMPLE

TASK

First-grade children are asked to arrange four pictures of trees in the order of the seasons by pasting them in four boxes and printing the name of each season in the box.

SCORING RUBRIC

2 points: Student arranges the pictures in the right order, beginning with any season.
1 point: Student begins the task but does not complete arrangement.
0 points: Student does not respond appropriately.

Task and scoring guide adapted from part of a Utah State Office of Education set of assessment tasks called *Weathercaster's Helper* for first-grade students (Regional Educational Laboratories, 1998).

Scoring rubrics for hands-on tasks may include multiple dimensions, each of which focuses on a particular aspect of the process of carrying out the task. For example, in an elementary school science task used by Shavelson, Baxter, and Pine (1991) and Shavelson, Baxter, and Gao (1993), students were asked to determine which of several paper towels absorbed the most water. The scoring rubric records the method used to get the towel wet, the saturation of each towel, the procedure used to measure the amount of water absorbed, the care in measurement, and the accuracy of the result.

Rating scales provide a flexible way of converting information about one or more characteristics of a performance (e.g., overall quality, adequacy of measurement, and appropriateness of summary of results). Typically, a rating scale consists of a set of characteristics or qualities to be judged and some type of scale for indicating the degree to which each attribute is present. The rating form itself is merely a reporting device. Its value in appraising the learning and development of students depends largely on the care with which it is prepared and the appropriateness with which it is used. As with other assessment instruments, it should be constructed in accordance with the learning outcomes to be assessed, and its use should be confined to those areas in which there is a sufficient opportunity to make the necessary observations. If these two principles are properly applied, a rating scale will serve several important assessment functions: (a) It will direct observation

toward specific aspects of performance, (b) it will provide a common frame of reference for rating the performance of all students on the same set of characteristics, and (c) it will provide a convenient method for recording the observer's judgments.

Types of Rating Scales

Rating scales may take many forms, but most of them belong to one of the types described next. Each type is illustrated by using two dimensions from a scale for rating contributions to class discussion.

Numerical Rating Scale. One of the simplest types of rating scales is that in which the rater checks or circles a number to indicate the degree to which a characteristic is present. Typically, each of a series of numbers is given a verbal description that remains constant from one characteristic to another. In some cases, it is merely indicated that the largest number is high, one is low, and the other numbers represent intermediate values.

The numerical rating scale is useful when the characteristics or qualities to be rated can be classified into a limited number of categories and there is general agreement concerning the category represented by each number. As commonly used, however, the numbers are only vaguely defined, so the interpretation and use of the scale vary.

EXAMPLE

Directions: Indicate the degree to which this student contributes to a group problem-solving task by circling the appropriate number. The numbers represent the following values: 4—consistently appropriate and effective; 3—generally appropriate and effective; 2—needs improvement, may wander from topic; and 1—unsatisfactory (disruptive or off topic).

1. To what extent does the student participate in group discussions?

   1 2 3 4

2. To what extent are the comments related to the topic under discussion?

   1 2 3 4

Graphic Rating Scale. The distinguishing feature of the graphic rating scale is that each characteristic is followed by a horizontal line. The rating is made by placing a check on the line. A set of categories identifies specific positions along the line, but the rater is free to check between these points.

EXAMPLE

Directions: Indicate the degree to which this student contributes to a group problem-solving task by placing an *X* anywhere along the horizontal line under each item.

1. To what extent does the student participate in group discussion?

   never seldom occasionally frequently always

2. To what extent are the comments related to the topic under discussion?

   never seldom occasionally frequently always

The scale shown in this example uses the same set of categories for each characteristic and is commonly referred to as a **constant-alternatives scale**. When these categories vary from one characteristic to another, the scale is called, quite logically, a **changing-alternatives scale**.

Although the line in the graphic rating scale makes it possible to rate at intermediate points, using single words to identify the categories has no great advantage over the use of numbers. There is little agreement among raters concerning the meaning of such terms as ***seldom, occasionally,*** and ***frequently.*** What is needed are descriptions of performances that indicate more specifically how students behave who possess various degrees of the characteristic being rated.

Descriptive Graphic Rating Scale. The descriptive graphic rating scale uses descriptive phrases to identify the points on a graphic scale. The descriptions are thumbnail sketches of how students behave at different steps along the scale. In some scales, only the center and end positions are defined. In others, a descriptive phrase is placed beneath each point. A space for comments is also frequently provided to enable the rater to clarify the rating.

EXAMPLE *Directions:* Make your ratings on each of the following characteristics by placing an *X* anywhere along the horizontal line under each item. In the space for comments, include anything that helps clarify your rating.

1. To what extent does the student participate in group discussions?

Never participates; quiet, passive	Participates as much as other group members	Participates more than any other group member

Comment:

2. To what extent are the comments related to the topic under discussion?

Comments ramble, distracts from topic	Comments usually pertinent, occasionally wanders from topic	Comments are always related to topic

Comment:

The descriptive graphic rating scale is generally the most satisfactory for school use. It explains to both the teacher and the student the types of performance that represent different degrees of progress toward desired learning outcomes. In well-written rubrics, the top level of description actually is the desired learning outcome or at least communicates what good work is intended to look like. The more specific performance descriptions also

contribute to greater objectivity and accuracy during the rating process. To aid scoring, numbers also may be added to each position on the scale.

Uses of Rating Scales

Rating scales can be used to assess a wide variety of learning outcomes and aspects of development. As a matter of convenience, these uses may be classified into two assessment areas: (1) process or procedure, and (2) product.

Process or Procedure Assessment. In many areas, achievement is expressed specifically through the student's performance. Examples include the ability to give a speech, manipulate laboratory equipment, work effectively in a group, sing, play a musical instrument, and perform various physical feats. Such activities do not result in a product that can be assessed, and short-answer or fixed-response tests are generally inadequate. Consequently, the process or procedures used in the performance itself must be observed and judged.

Rating scales are especially useful in assessing process or procedures because they focus on the same aspects of performance in all students and have a common scale on which to record our judgments. If the rating form has been prepared in terms of specific learning outcomes, it also serves as an excellent teaching device. The dimensions and behavior descriptions used in the scale show the student the type of performance desired.

Two items from a typical rating scale for assessing a speech are presented in Figure 11.5. The first part of the form is devoted to the content of the speech and how well it is organized. The second part is concerned with aspects of delivery, such as gestures, posture, appearance, eye contact, voice, and enunciation. In developing such a scale, a teacher must, of course, include those characteristics that are most appropriate for the type of speaking ability to be assessed and for the age level of the student to be judged.

Product Assessment. When student performance results in some type of product, it is frequently more desirable to judge the product than the process or procedures. The ability to write a theme, for example, is best assessed by judging the quality of the theme itself. Little is to be learned by observing the student's performance. In some areas, however, such as word processing, conducting work in the laboratory, and woodworking, it might be most desirable to rate procedures during the early phase of learning and products later, after the basic skills have been mastered. In any event, product rating can provide assessment information in many areas. In addition to those already mentioned, it is useful in assessing such things as handwriting, drawings, maps, graphs, notebooks, term papers, book reports, results of laboratory experiments, and objects made in vocational courses.

A rating scale serves somewhat the same purpose in product assessment than it does in process assessment. It helps us judge the products of all students in terms of the same characteristics, and it emphasizes to the students those qualities desired in a superior product.

Speech Rating Scale

Directions: Rate the student's speaking ability by placing an *X* anywhere along the horizontal line under each characteristic. In the space for comments, include anything that helps clarify your rating or further describes the student's speech behavior.

A. Content and Organization

1. Opening remarks

Inappropriate; distract from speech topic.	Commonplace; no particular contribution to the speech.	Arouse interest; direct attention to speech topic.

Comment:

B. Delivery

2. Gestures

Movements are monotonous or distracting.	Generally effective; some distracting mannerisms.	Natural, expressive movements that emphasize speech.

Comment:

Figure 11.5
Sample items from speech rating scale

Common Errors in Rating

Certain types of errors occur so often in ratings that special efforts are needed to counteract them. These include (a) personal bias, (b) halo effect, and (c) logical errors.

Personal bias errors occur when there is a general tendency to rate all individuals at approximately the same position on the scale. Some raters tend to use the high end of the scale only, which is referred to as the **generosity error.** Occurring less frequently (but persistently for some raters) is the **severity error,** in which the lower end of the scale is favored. A third type of constant response is shown by the rater who avoids both extremes of the scale and tends to rate everyone as average. This is called the **central tendency error.**

It also occurs much less often than the generosity error, but it tends to be a fixed-response style for some raters.

The tendency of a rater to favor a certain position on the scale has two undesirable results. First, it puts in doubt a single rating of an individual. A high or low rating might reflect the personal outlook of the rater rather than the actual performance or personal characteristics of the person rated. Second, favoring a certain position on the scale limits the range of any individual's ratings. Therefore, even if we make allowances for a teacher's general tendency to rate students high, the ratings for different students may be so close together that they fail to provide reliable discriminations.

The **halo effect** is an error that occurs when a rater's general impression of a person influences the rating of individual characteristics. If the rater has a favorable attitude toward the person being rated, then there will be a tendency to give high ratings on all traits; but if the rater's attitude is unfavorable, the ratings will be low. This differs from the generosity and severity errors, in which the rater tends to rate everyone high or everyone low.

Because the halo effect causes a student to receive similar ratings on all characteristics, it tends to obscure strengths and weaknesses on different traits. This obviously limits the value of the ratings.

Teachers need to guard against the possibility that their ratings might be distorted because of preconceptions based on inappropriate factors such as gender, race, ethnicity, and social background. Halo effects leading to lowered ratings of all performances of some students as the result of such preconceptions are of particular concern. Concealing the identity of the student where feasible when rating products of performance is one good safeguard against halo effects. Awareness of our own personal preferences and prejudices is also important.

A logical error results when two characteristics are rated as more alike or less alike than they actually are because of the rater's beliefs concerning their relationship. In rating achievement, for example, teachers tend to overrate the achievement of students identified by aptitude tests as gifted because they expect achievement and giftedness to go together. Similarly, teachers who hold the common but false belief that gifted students have poor social adjustment will tend to underrate them on social characteristics. These errors result not from biases toward certain students or certain positions on the rating scale but from the rater's assumption of a more direct relationship among traits than actually exists.

The various types of errors that appear in ratings are rather disconcerting to the classroom teacher who must depend on rating scales for assessing certain aspects of learning and development. Fortunately, however, the errors can be markedly reduced by proper design and use.

Principles of Effective Rating

The improvement of ratings requires careful attention to selection of the characteristics to be rated, design of the rating form, and conditions under which the ratings are obtained. The following principles summarize the most important considerations in these areas. Because the descriptive graphic rating scale is the most generally useful form for

school purposes, the principles are directed toward the construction and use of this type of rating scale.

1. Characteristics should be educationally significant. Rating scales, like other assessment instruments, must be in harmony with the school's objectives and desired learning outcomes. Thus, when constructing or selecting a rating scale, the best guide for determining what characteristics are most significant is the list of intended learning outcomes. When these have been clearly stated in performance terms, it is often simply a matter of selecting those that can be most effectively assessed by ratings and then modifying the statements to fit the rating format (see the "Guidelines" box).

GUIDELINES

Preparing Rating Scales

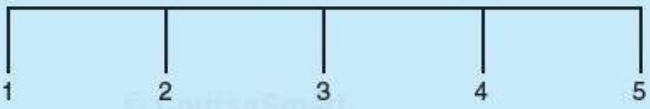
The same basic principle guiding the construction of test items should be followed in preparing rating scales. That is, the instrument should be designed to measure the student performance described in the instructional objectives. Let us assume, for example, that a science teacher has listed the following outcomes as evidence of skill in one phase of laboratory performance.

Demonstrates Effective Use of Laboratory Equipment

1. Selects proper equipment for a given experiment.
2. Sets up equipment quickly and correctly.
3. Manipulates equipment as needed during the experiment.
4. Measures accurately with each measuring device.
5. Follows safety rules when using equipment.
6. Cleans and returns equipment to its proper place.
7. Interprets the results of the experiment appropriately.
8. Integrates results with other knowledge in drawing conclusions.

This list of intended outcomes can then serve as the basis for preparing a rating scale to assess skill in using laboratory equipment. Each item in the list becomes an item in the rating form by simply adding some basis for recording degrees of effectiveness, as follows:

Selecting Laboratory Equipment



Cannot select equipment without help	Inconsistent in selecting proper equipment	Consistently selects proper equipment

The same procedure is followed when rating an educational product (e.g., theme, graph, painting, or shop and home economics projects). The characteristics of a good product are listed, and these then become the items in the rating scale. The instrument itself is simply a convenient form for recording observations and judgments concerning the extent to which students are meeting the criteria specified in the objectives.

2. Identify the learning outcomes that the task is intended to assess. The intent of the assessment is critical for determining those characteristics of performance that should determine the ratings. Clear identification of the learning outcomes helps establish priorities for rating, distinguish levels of performance in terms of learning outcomes, and reduce dependence on factors that are irrelevant to the intent of the assessment. When there are multiple learning outcomes associated with the task, separate ratings corresponding to each outcome may be desirable and can enhance the value of the formative feedback that is provided to students.

3. Characteristics should be directly observable. There are two considerations involved in direct observation. First, the characteristics should be limited to those that occur in school situations so the teacher has an opportunity to observe them. Second, they should be characteristics that are clearly visible to an observer. Overt behaviors, such as participation in classroom discussion, clear enunciation, and use of facts to support an argument, can be readily observed and reliably rated. However, less tangible types of behavior, such as interest in history, attitude toward literature, and amount of effort expended in library research, tend to be unreliably rated because their presence must be inferred from outward signs that are indefinite, variable, and easily faked. When possible, we should confine our ratings to those characteristics that can be observed and judged directly.

4. Characteristics and points on the scale should be clearly defined. Many rating errors arise from the use of vague characterizations and inadequate identification of the scale points. The brief descriptions used with the descriptive graphic rating scale help overcome this weakness. They explain both the points on the scale and each characteristic being rated. When it is infeasible or inconvenient to use a descriptive scale, as on the back of a school report card, a separate sheet of instructions can be used to provide the desired descriptions.

5. Select the type of scoring rubric that is most appropriate for the task and the purpose of the assessment. With a holistic rubric, each performance is given a single rating or score, usually on a scale with 4 to 6 points, based on an overall judgment of the quality of the performance in comparison to the criteria specified in the scoring rubric. Holistic rubrics are efficient and translate easily into grades. As already noted, however, analytic scoring rubrics have more diagnostic value because they focus attention on those aspects of performance where improvement is needed. For analytic scores to be of diagnostic value, the characteristics or dimensions being rated must be sufficiently distinct to allow each to be reliably rated and not simply be redundant reflections of the same global impression of the performance.

6. Between three and seven rating positions should be provided. The exact number of points to be designated on a particular scale is determined largely by the judgments to be made. In areas permitting only crude judgments, fewer scale positions are needed. There is usually no advantage in going beyond the 7-point scale. Only rarely can we make finer discriminations than this, and we can provide for those few situations by allowing the rater to mark between points.

7. Rate performances of all students on one task before going on to the next one. The advantages of rating all performances on one task before starting another task parallel those described for scoring all answers to an essay question before going on to the next

question. It is easier to keep the scoring criteria clearly in mind and to apply them more uniformly when considering only a single task at a time than it is when going from task to task for each student. It also reduces the likelihood that judgments of performance on one task will be contaminated by judgments of a student's performance on a preceding task. When responses of a single student to several tasks are considered one after another, there is a strong tendency for the performance on early tasks to create an expectation for performance on later tasks. Those expectations can result in more lenient or more stringent ratings of performance than would otherwise be given.

By rating one task at a time for all students before going to the next task, it is also possible to change the order in which student performances are rated. Thus, a student is not rated first or last on all tasks or right after another student who has exceptionally good performance or exceptionally bad performance on all tasks.

8. **When possible, rate performances without knowledge of the student's name.** This suggestion is the same as the one given for scoring answers to essay questions. Obviously, it is not possible for all types of performance (e.g., an oral presentation), but it is good practice when possible. It is a practice that enhances the fairness of ratings because it reduces the chances that ratings will be influenced by a halo effect rather than only by the actual performance of a student.

9. **When results from a performance assessment are likely to have long-term consequences for students, ratings from several observers should be combined.** The pooled ratings of several teachers will generally yield a more reliable description of student performance than that obtained from any one teacher. In averaging ratings, the personal biases of individual raters tend to cancel out one another, but there is still a need to be alert for biases that may be shared due to similarity of background and experiences of teachers doing the rating.

CHECKLISTS

A checklist is similar in appearance and use to the rating scale. The basic difference between them is in the type of judgment needed. On a rating scale, one can indicate the degree to which a characteristic is present or the frequency with which a behavior occurs. The checklist, on the other hand, calls for a simple yes–no judgment. It is basically a method of recording whether a characteristic is present or absent or whether an action was or was not taken. Obviously, a checklist should not be used when degree or frequency of occurrence is an important aspect of the appraisal.

The checklist is especially useful at the primary level, where much of the classroom assessment depends on observation rather than testing. A simple checklist for assessing the mastery of mathematics skills at the beginning primary level is shown in Figure 11.6. If the intended learning outcomes are stated as specifically as this for each learning area, a checklist can be prepared by simply adding a place to check yes or no. As with the rating scales, the stated learning outcomes specify the performance to be assessed, and the checklist is merely a convenient means of recording judgments.

Checklists are also useful in assessing those performance skills that can be divided into a series of specific actions. An example of such a checklist for the proper application

Mathematics Skills Checklist

Primary Level

Directions: Circle YES or NO to indicate whether skill has been demonstrated.

YES	NO	1. Identifies numerals 0 to 10.
YES	NO	2. Counts to 10.
YES	NO	3. Groups objects into sets of 1 to 10.
YES	NO	4. Identifies basic geometric shapes (circle, square, rectangle, triangle).
YES	NO	5. Identifies coins (penny, nickel, dime).
YES	NO	6. Compares objects and identifies bigger–smaller, longer–shorter, heavier–lighter.
YES	NO	7. States ordinals for a series of 10 objects (1st, 2nd, 3rd, etc.).
YES	NO	8. Copies numerals 1 to 10.
YES	NO	9. Tells time to the half hour.
YES	NO	10. Identifies one-half of an area.

Figure 11.6

Checklist for evaluating student's mastery of begining skills in mathematics

of varnish is shown in Figure 11.7. The performance has been subdivided into a series of observable steps, and the observer simply checks whether each step was satisfactorily completed. The checklist in Figure 11.7 includes mostly those actions that are desired in a good performance. In some cases, it may be useful to add those actions that represent common errors so that they can be checked if they occur. In Figure 11.7, for example, we might add after Item 4, "Does ***not*** stir varnish before using." Because stirring paint is a necessary step when painting, some students might incorrectly carry over this action when using varnish. If the checklist is to be used by students, the incorrect actions should, of course, be clearly identified as such.

The following steps summarize the development of a checklist for assessing a procedure consisting of a series of sequential steps.

1. Identify each of the specific actions desired in the performance.
2. Add to the list those actions that represent common errors (if they are useful in the assessment, are limited in number, and can be clearly stated).
3. Arrange the desired actions (and likely errors, if used) in the approximate order in which they are expected to occur.
4. Provide a simple procedure for checking each action as it occurs (or for numbering the actions in sequence, if appropriate).

In addition to its use in assessment of process, the checklist can also be used to assess products. For this purpose, the form usually contains a list of characteristics that the finished product should possess. In assessing the product, the teacher simply checks whether each characteristic is present or absent. Before using a checklist for product assessment, you should decide whether the quality of the product can be adequately described by merely noting the presence or absence of each characteristic. If quality is

Directions: On the space in front of each item, place a plus (+) sign if performance was satisfactory, place a minus (−) sign if it was unsatisfactory.

_____ 1. Sands and prepares surface properly.
_____ 2. Wipes dust from surface with appropriate cloth.
_____ 3. Selects appropriate brush.
_____ 4. Selects varnish and checks varnish flow.
_____ 5. Pours needed amount of varnish into clean container.
_____ 6. Puts brush properly into varnish (1/3 of bristle length).
_____ 7. Wipes excess varnish from brush on inside edge of container.
_____ 8. Applies varnish to surface with smooth strokes.
_____ 9. Works from center of surface toward the edges.
_____ 10. Brushes with the grain of the wood.
_____ 11. Uses light strokes to smooth the varnish.
_____ 12. Checks surface for completeness.
_____ 13. Cleans brush with appropriate cleaner.
_____ 14. Does *not* pour excess varnish back into can.
_____ 15. Cleans work area.

Figure 11.7
Checklist for evaluating the proper application of varnish

Source: N. E. Gronlund, *Stating Objectives for Classroom Instruction,* 3rd ed. Copyright 1985, Prentice Hall, New Jersey. Used by permission.

more precisely indicated by noting the degree to which each characteristic is present, a rating scale should be used instead of a checklist.

In the area of personal–social development, the checklist can be a convenient method of recording evidence of growth toward specific learning outcomes. Typically, the form lists the behaviors that have been identified as representative of the outcomes to be assessed. In the area of work habits, for example, a primary teacher might list the following behaviors (to be marked yes or no):

- Follows directions
- Seeks help when needed
- Works cooperatively with others
- Waits turn in using materials
- Shares materials with others
- Tries new activities
- Completes started tasks
- Returns equipment to proper place
- Cleans work space

Although such items can be used in checklist form if only a crude appraisal is desired, they can also be used in rating scale form by recording the frequency of occurrence (e.g., always, sometimes, never).

Although we have described the individual use of checklists, rating scales, and anecdotal records (see Chapter 13), they are often used in combination when assessing student performance (see Table 11.2).

Table 11.2
Combining techniques to assess laboratory performance in science

Types of Proficiency	*Examples of Performance to Be Assessed*	*Assessment Techniques*
Knowledge of experimental procedures	Describes relevant procedures Identifies equipment and uses Criticizes defective experiments	Paper-and-pencil testing Laboratory identification tests
Skill in designing an experiment	Plans and designs an experiment to be performed	Performance assessment with focus on product (checklist)
Skill in conducting the experiment	Selects equipment Sets up equipment Conducts experiment	Performance assessment with focus on process (rating scale)
Skill in observing and recording	Describes procedures used Reports proper measurements Organizes and records results	Performance assessment (analysis of report)
Skill in interpreting results	Identifies significant relationships Identifies weaknesses in data States valid conclusions	Performance assessment and oral questioning
Work habits	Manipulates equipment effectively Completes work promptly Cleans work space	Performance assessment with focus on process (checklist)

STUDENT PARTICIPATION IN RATING

In this chapter, we have limited our discussion to rating scales and checklists used by the teacher. We purposely omitted those checklists and rating scales used as self-report techniques by students because these will be considered in the following chapter. Before closing our discussion here, however, we should point out that most of the devices used for recording the teacher's observations also can be used by students to judge their own progress. From an instructional standpoint, it is often useful to have students rate themselves (or their products) and then compare the ratings with those of the teacher. If this comparison is made during an individual conference, the teacher can explore with each student the reasons for the ratings and discuss any marked discrepancies between the two sets.

Self-rating by a student and a follow-up conference with the teacher can have many benefits. It should help the student (a) understand better the instructional objectives, (b) recognize the progress being made toward the objectives, (c) diagnose more effectively particular strengths and weaknesses, and (d) develop increased skill in self-assessment. Of special value to the teacher is the additional insight gained.

Student participation need not be limited to the use of the assessment instruments. It is also useful to have students help develop the instruments. Through class discussion, for example, they can help identify the qualities desired in a good speech or a well-written

report. A list of these suggestions can then be used as a basis for constructing a rating scale or checklist. Involving students in the development of assessment devices has special instructional values. First, it directs learning by causing the students to think more carefully about the qualities to strive for in a performance or product. Second, it has a motivating effect because students tend to put forth most effort when working toward goals they have helped define.

SUMMARY

Performance tasks provide a means of assessing a variety of student skills that cannot be measured by objective tests. To name just a few of the possibilities in addition to written responses, the performances may include oral communication; the construction of models, graphs, diagrams, or maps; or the use of tools and equipment (computers, or scientific or musical instruments). Unlike objective items, both the **process** and the **product** resulting from the performance can be assessed. Because they are time consuming both for students to do and for teachers to rate, the emphasis on performance assessment should be on measuring complex achievement that cannot be measured well by objective tests.

Restricted-response tasks are more structured and require less time to administer than extended-response tasks. These features facilitate reliability and wider coverage of a content domain. Extended-response tasks are best suited to the measurement of more complex learning outcomes, such as gathering, organizing, synthesizing, evaluating, and presenting information.

Extended performance tasks underscore the importance attached to effective performance and provide an effective means of measuring significant learning outcomes. They are the only feasible approach for measuring some important learning outcomes, they allow for the assessment of process as well as product, and their emphasis on the engagement of students in the active construction of meaning is consistent with modern learning theory. Their limitations are due mainly to the unreliability of judgmental ratings and to the time-consuming nature of the tasks and rating. Careful attention to rating criteria is critical for minimizing the unreliability due to scoring. Because of the limited generalizability of performance across tasks designed to measure the same or similar learning outcomes, it is important to base decisions on evidence accumulated from several tasks.

Rating methods are a systematic procedure for obtaining and recording the observers' judgments. Of the several types of rating scales available, the descriptive graphic scale seems to be the best for school use. In rating procedures, products, and various aspects of personal–social development, certain types of errors commonly occur. These include personal bias, halo effect, and logical errors. The control of such errors is a major consideration in constructing and using rating scales. Effective ratings result when we (a) select educationally significant characteristics, (b) identify the learning outcomes that the task is intended to assess, (c) limit ratings to directly observable behavior, (d) define clearly the characteristics and the points on the scale, (e) select the most appropriate rating procedure, (f) limit the number of points on the scale, (g) rate performances of all students on one task before going on to the next ones, (h) rate performances without

knowledge of the student's name when possible, and (i) combine ratings from several raters when results may have long-term consequences for students.

Checklists perform somewhat the same functions as rating scales. They are used in assessing both process and products where assessment is limited to a simple present–absent judgment.

Having students help construct and use rating devices has special values from the standpoint of learning and aids in the development of self-assessment skills.

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 performance-based assessment tasks. For each learning outcome, construct two tasks.
2. What factors should be considered in deciding whether extended performance assessment tasks are to be included in a classroom assessment? Which of the factors are most important?
3. Describe how performance assessments might be used to facilitate learning. What types of learning are most likely to be enhanced?
4. Construct a rating scale for one of the following that would be useful for assessing the effectiveness of the performance.
   a. Giving an oral report
   b. Working in the laboratory
   c. Participating in group work
   d. Playing some type of game
   e. Demonstrating a skill
5. Construct a rating scale or checklist for one of the following that would be useful for assessing the product.
   a. Constructing a map, chart, or graph
   b. Writing a personal or business letter
   c. Writing a theme, poem, or short story
   d. Making a drawing or painting
   e. Making a product in home economics
   f. Making a product in industrial education
6. Prepare a checklist for assessing the ability to drive an automobile. Would a rating scale be better for this purpose? What are the relative advantages of each?
7. List some of the areas of assessment in which product scales might be used for rating.

REFERENCES

Baker, E. L., Aschbacher, P. R., Niemi, D., & Sato, E. (1992). *CRESST performance assessment models: Assessing content area explanations.* Los Angeles: University of California Center for Research on Evaluation, Standards, and Student Testing.

Mathematical Sciences Education Board. (1993). *Measuring up: Prototypes for mathematics assessment.* Washington, DC: National Academy Press. See pages 141–155 for a discussion of this game and related assessment questions.

O'Sullivan, C. Y., Reese, C. M., & Mazzeo, J. (1997). *NAEP 1996 Science Report Card for the Nation and the States.* Washington, DC: National Center for Education Statistics. Available: http://www.ed.gov/NCES/naep

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

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30,* 215–232.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4,* 347–362.

Stiggins, R. J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practice, 6*(3), 33–42. Part of an instructional series of the National Council on Measurement in Education. It presents helpful guidelines for the construction of performance assessments.

FURTHER READING

Educational Testing Service. (1993). *Performance assessment sampler: A workbook.* Princeton, NJ: Educational Testing Service. This workbook presents examples of performance assessments in various subjects, with examples of student responses and scores assigned.

Gronlund, N. E. (2005). *Assessment of student achievement* (8th ed.). Boston: Allyn & Bacon. Chapter 7, "Traditional Performance Assessments of Skills and Products," and Chapter 8, "Expanded Performance Assessments," discuss the construction and use of various types of performance assessments.

Hart, D. (1994). *Authentic assessment: A handbook for educators.* Menlo Park, CA: Addison-Wesley. Provides a variety of examples of performance-based assessments and arguments for the importance of this approach to assessment.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development. In addition to providing examples of performance assessments and guidelines for rating, the book presents a model for linking assessment and instruction.

Lane, S., & Stone, C. (2006). Performance assessment. In R. L Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 387–433). Westport, CT: American Council on Education/ Praeger. Provides a detailed discussion of the types and uses of performance assessments and examples of holistic and analytic scoring rubrics.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., & Phillips, G. W. (1993). *NAEP 1992: Mathematics Report Card for the Nation and the States* (Report No. 23-STO2). Washington, DC: U.S. Department of Education.

Welch, C. (2006). Item and prompt development in performance testing. In S. M Downing & T. M Haladyna (Eds.). *Handbook of Test development* (Chapter 13, pp. 303-327). Mahway, NJ: Lawrence Erlbaum. Provides guidance for improving the quality of performance assessments and gives examples of scoring rubrics.