

Best Practices for Designing and Grading Exams

Adapted from M.E. Piontek (2008)
Center for Research on Learning and Teaching

The most obvious function of assessment methods (such as exams, quizzes, papers, and presentations) is to enable instructors to make judgments about the quality of student learning (i.e., assign grades). However, the method of assessment also can have a direct impact on the quality of student learning. Students assume that the focus of exams and assignments reflects the educational goals most valued by an instructor, and they direct their learning and studying accordingly (McKeechie & Svinicki, 2006). General grading systems can have an impact as well. For example, a strict bell curve (i.e., *norm-reference grading*) has the potential to dampen motivation and cooperation in a classroom, while a system that strictly rewards proficiency (i.e., *criterion-referenced grading*) could be perceived as contributing to grade inflation. Given the importance of assessment for both faculty and student interactions about learning, how can instructors develop exams that provide useful and relevant data about their students' learning and also direct students to spend their time on the important aspects of a course or course unit? How do grading practices further influence this process?

Guidelines for Designing Valid and Reliable Exams

Ideally, effective exams have four characteristics. They are:

- *Valid* (providing useful information about the concepts they were designed to test),
- *Reliable* (allowing consistent measurement and discriminating between different levels of performance),
- *Recognizable* (instruction has prepared students for the assessment), and
- *Realistic* (concerning time and effort required to complete the assignment) (Svinicki, 1999).

Most importantly, exams and assignments should *focus on the most important content and behaviors* emphasized during the course (or particular section of the course). What are the primary ideas, issues, and skills you hope students learn during a particular course/unit/module? These are the *learning outcomes* you wish to measure. For example, if your learning outcome involves memorization, then you should assess for memorization or classification; if you hope students will develop problem-solving capacities, your exams should focus on assessing students' application and analysis skills. As a general rule, assessments that focus too heavily on details (e.g., isolated facts, figures, etc.) "will probably lead to better student retention of the footnotes at the cost of the main points" (Halpern & Hakel, 2003, p. 40). As noted in Table 1, each type of exam item may be better suited to measuring some learning outcomes than others, and each has its advantages and disadvantages in terms of ease of design, implementation, and scoring.

Table 1: Advantages and Disadvantages of Commonly Used Types of Achievement Test Items

Type of Item	Advantages	Disadvantages
True-False	Many items can be administered in a relatively short time. Moderately easy to write; easily scored.	Limited primarily to testing knowledge of information. Easy to guess correctly on many items, even if material has not been mastered.
Multiple-Choice	Can be used to assess broad range of content in a brief period. Skillfully written items can measure higher order cognitive skills. Can be scored quickly.	Difficult and time consuming to write good items. Possible to assess higher order cognitive skills, but most items assess only knowledge. Some correct answers can be guesses.
Matching	Items can be written quickly. A broad range of content can be assessed. Scoring can be done efficiently.	Higher order cognitive skills are difficult to assess.
Short Answer or Completion	Many can be administered in a brief amount of time. Relatively efficient to score. Moderately easy to write.	Difficult to identify defensible criteria for correct answers. Limited to questions that can be answered or completed in very few words.
Essay	Can be used to measure higher order cognitive skills. Relatively easy to write questions. Difficult for respondent to get correct answer by guessing.	Time consuming to administer and score. Difficult to identify reliable criteria for scoring. Only a limited range of content can be sampled during any one testing period.

Adapted from Table 10.1 of Worthen, et al., 1993, p. 261.

General Guidelines for Developing Multiple-Choice and Essay Questions

The following sections highlight general guidelines for developing multiple-choice and essay questions, which are often used in college-level assessment because they readily lend themselves to measuring higher order thinking skills (e.g., application, justification, inference, analysis and evaluation). Yet instructors often struggle to create, implement, and score these types of questions (McMillan, 2001; Worthen, et al., 1993).

Multiple-choice questions have a number of advantages. First, they can measure various kinds of knowledge, including students' understanding of terminology, facts, principles, methods, and procedures, as well as their ability to apply, interpret, and justify. When carefully designed, multiple-choice items also can assess higher-order thinking skills.

Multiple-choice questions are less ambiguous than short-answer items, thereby providing a more focused assessment of student knowledge. Multiple-choice items are superior to true-false items in several ways: on true-false items, students can receive credit for knowing that a statement is incorrect, without knowing what is correct. Multiple-choice items offer greater reliability than true-false items as the opportunity for guessing is reduced with the larger number of options. Finally, an instructor can diagnose misunderstanding by analyzing the incorrect options chosen by students.

A disadvantage of multiple-choice items is that they require developing incorrect, yet plausible, options that can be difficult to create. In addition, multiple-choice questions do not allow instructors to measure students' ability to organize and present ideas. Finally, because it is much easier to create multiple-choice items that test recall and recognition rather than higher order thinking, multiple-choice exams run the risk of not assessing the deep learning that many instructors consider important (Greenland & Linn, 1990; McMillan, 2001).

Guidelines for writing multiple-choice items include advice about stems, correct answers, and distractors (McMillan, 2001, p. 150; Piontek, 2008):

Stems pose the problem or question.

- Is the stem stated as clearly, directly, and simply as possible?
- Is the problem described fully in the stem?
- Is the stem stated positively, to avoid the possibility that students will overlook terms like “no,” “not,” or “least”?
- Does the stem provide only information relevant to the problem?

Possible responses include the correct answer and *distractors*, or the incorrect choices. Multiple-choice questions usually have at least three distractors.

- Are the distractors plausible to students who do not know the correct answer?
- Is there only one correct answer?
- Are all the possible answers parallel with respect to grammatical structure, length, and complexity?
- Are the options short?
- Are complex options avoided? Are options placed in logical order?
- Are correct answers spread equally among all the choices? (For example, is answer “A” correct about the same number of times as options “B” or “C” or “D”)?

An example of good multiple-choice questions that assess higher-order thinking skills is the following test question from pharmacy (Park, 2008):

Patient WC was admitted for third-degree burns over 75% of his body. The attending physician asks you to start this patient on antibiotic therapy. Which one of the following is the best reason why WC would need antibiotic prophylaxis?

- a. His burn injuries have broken down the innate immunity that prevents microbial invasion.
- b. His injuries have inhibited his cellular immunity.
- c. His injuries have impaired antibody production.
- d. His injuries have induced the bone marrow, thus activated immune system

A second question builds on the first by describing the patient's labs two days later, asking the students to develop an explanation for the subsequent lab results. (See Piontek, 2008 for the full question.)

Essay questions can tap complex thinking by requiring students to organize and integrate information, interpret information, construct arguments, give explanations, evaluate the merit of ideas, and carry out other types of reasoning (Cashin, 1987; Gronlund & Linn, 1990; McMillan, 2001; Thorndike, 1997; Worthen, et al., 1993). *Restricted response* essay questions are good for assessing basic knowledge and understanding and generally require a brief written response (e.g., “State

two hypotheses about why birds migrate. Summarize the evidence supporting each hypothesis” [Worthen, et al., 1993, p. 277].) *Extended response* essay items allow students to construct a variety of strategies, processes, interpretations and explanations for a question, such as the following:

The framers of the Constitution strove to create an effective national government that balanced the tension between majority rule and the rights of minorities. What aspects of American politics favor majority rule? What aspects protect the rights of those not in the majority? Drawing upon material from your readings and the lectures, did the framers successfully balance this tension? Why or why not? (Shipan, 2008).

In addition to measuring complex thinking and reasoning, advantages of essays include the potential for motivating better study habits and providing the students flexibility in their responses. Instructors can evaluate how well students are able to communicate their reasoning with essay items, and they are usually less time consuming to construct than multiple-choice items that measure reasoning.

The major disadvantages of essays include the amount of time instructors must devote to reading and scoring student responses, and the importance of developing and using carefully constructed criteria/rubrics to insure reliability of scoring. Essays can assess only a limited amount of content in one testing period/exam due to the length of time required for students to respond to each essay item. As a result, essays do not provide a good sampling of content knowledge across a curriculum (Gronlund & Linn, 1990; McMillan, 2001).

Guidelines for writing essay questions include the following (Gronlund & Linn, 1990; McMillan, 2001; Worthen, et al., 1993):

- Restrict the use of essay questions to educational outcomes that are difficult to measure using other formats. For example, to test recall knowledge, true-false, fill-in-the-blank, or multiple-choice questions are better measures.
- Identify the specific skills and knowledge that will be assessed. Piontek (2008, available online: <http://www.crlt.umich.edu/resources/occasional>) gives several examples of question stems that assess different types of reasoning skills, such as:
 - *Generalizations*: State a set of principles that can explain the following events.
 - *Synthesis*: Write a well-organized report that shows...
 - *Evaluation*: Describe the strengths and weaknesses of...
- Write the question clearly so that students do not feel that they are guessing at “what the instructor wants me to do.”
- Indicate the amount of time and effort students should spend on each essay item.
- Avoid giving students options for which essay questions they should answer. This choice decreases the validity and reliability of the test because each student is essentially taking a different exam.
- Consider using several narrowly focused questions (rather than one broad question) that elicit different aspects of students’ skills and knowledge.
- Make sure there is enough time to answer the questions.

Guidelines for scoring essay questions include the following (Gronlund & Linn, 1990; McMillan, 2001; Wiggins, 1998; Worthen, et al., 1993; *Writing and grading essay questions*, 1990):

- Outline what constitutes an expected answer.
- Select an appropriate scoring method based on the criteria. A *rubric* is a scoring key that indicates the criteria for scoring and the amount of points to be assigned for each criterion. A sample rubric for a take-home history exam question might look like the following:

Criteria	Level of performance		
	0 points	1 points	2 points
Number of references to class reading sources	0-2 references	3-5 references	6+ references
Historical accuracy	Lots of inaccuracies	Few inaccuracies	No apparent inaccuracies
Historical Argument	No argument made; little evidence for argument	Argument is vague and unevenly supported by evidence	Argument is clear and well-supported by evidence
Proof reading	Many grammar and spelling errors	Few (1-2) grammar or spelling errors	No grammar or spelling errors
Total Points (out of 8 possible):			

For other examples of rubrics, see Piontek (2008, available online: <http://www.crlt.umich.edu/resources/occasional>).

- Clarify the role of writing mechanics and other factors independent of the educational outcomes being measured. For example, how does grammar or use of scientific notation figure into your scoring criteria?
- Create anonymity for students' responses while scoring and create a random order in which tests are graded (e.g., shuffle the pile) to increase accuracy of the scoring.
- Use a systematic process for scoring each essay item. Assessment guidelines suggest scoring all answers for an individual essay question in one continuous process, rather than scoring all answers to all questions for an individual student. This system makes it easier to remember the criteria for scoring each answer.

You can also use these guidelines for scoring essay items to create grading processes and rubrics for students' papers, oral presentations, course projects, and websites. (For other grading strategies, see "Responding to Student Writing – Principles & Practices," p. 136 and "Responding to Student Writing – A Sample Commenting Protocol," p. 137).

References

- Cashin, W. E. (1987). *Improving essay tests*. Idea Paper, No. 17. Manhattan, KS: Center for Faculty Evaluation and Development, Kansas State University.
- Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan Publishing Company.
- Halpern, D. H., & Hakel, M. D. (2003). Applying the science of learning to the university and beyond. *Change*, 35(4), 37-41.
- McKeachie, W. J., & Svinicki, M. D. (2006). Assessing, testing, and evaluating: Grading is not the most important function. In McKeachie's *Teaching tips: Strategies, research, and theory for college and university teachers* (12th ed., pp. 74-86). Boston: Houghton Mifflin Company.
- McMillan, J. H. (2001). *Classroom assessment: Principles and practice for effective instruction*. Boston: Allyn and Bacon.
- Park, J. (2008, February 4). Personal communication. University of Michigan College of Pharmacy.
- Piontek, M. (2008). *Best practices for designing and grading exams*. CRLT Occasional Paper No. 24. Ann Arbor, MI. Center for Research on Learning and Teaching. Available: <http://www.crlt.umich.edu/resources/occasional>
- Shipan, C. (2008, February 4). Personal communication. University of Michigan Department of Political Science.
- Svinicki, M. D. (1999a). Evaluating and grading students. In *Teachers and students: A sourcebook for UT- Austin faculty* (pp. 1-14). Austin, TX: Center for Teaching Effectiveness, University of Texas at Austin.
- Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education*. Upper Saddle River, NJ: Prentice-Hall, Inc.
- Wiggins, G. P. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco: Jossey-Bass Publishers.
- Worthen, B. R., Borg, W. R., & White, K. R. (1993). *Measurement and evaluation in the schools*. New York: Longman.
- Writing and grading essay questions. (1990, September). *For Your Consideration*, No. 7. Chapel Hill, NC: Center for Teaching and Learning, University of North Carolina at Chapel Hill.