BEST PRACTICES FOR DESIGNING AND GRADING EXAMS

Mary E. Piontek

Introduction

The most obvious function of assessment methods such as exams, quizzes, papers, presentations, etc., is to enable instructors to make judgments about the quality of student learning (i.e., assign grades). However, the methods of assessment used by faculty can also 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 (McKeachie & Svinicki, 2006). 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?

Creating high quality educational assessments requires both art and science: the *art* of creatively engaging students in assessments that they view as fair and meaningful, and that produce relevant data about student achievement; and the *science* of assessment design, item writing, and grading procedures (Worthen, Borg, & White, 1993). This Occasional Paper provides an overview of the *science* of developing valid and reliable exams, especially multiple-choice and essay items. Additionally, the paper describes key issues related to grading: holistic and trait-analytic rubrics, and normative and criterion grading systems.

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, 1999a). The following six guidelines are designed to help you create such assessments:

1. *Focus on the most important content and behaviors* you have emphasized during the course (or particular section of the course). Start by asking yourself to identify the primary ideas, issues, and skills encountered by students during a particular course/unit/module. The distribution of items should reflect the relative emphasis you gave to

Mary E. Piontek is Evaluation Researcher at the Center for Research on Learning and Teaching (CRLT). She has a Ph.D. in Measurement, Research, and Evaluation.

CRLT Occasional Papers

Center for Research on Learning and Teaching

University of Michigan

No. 24

content coverage and skill development (Gronlund & Linn, 1990). If you teach for memorization then you should assess for memorization or classification; if you emphasize problem-solving strategies 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.) can have a negative impact on student learning. "Asking learners to recall particular pieces of the information they've been taught often leads to 'selective forgetting' of related information that they were not asked to recall." In fact, testing for "relatively unimportant points in the belief that 'testing for the footnotes' will enhance learning,…will probably lead to better student retention of the footnotes at the cost of the main points" (Halpern & Hakel, 2003, p. 40).

2. *Write clearly and simply*. Appropriate language level, sentence structure, and vocabulary are three essential characteristics of all assessment items. The items should be clear and succinct, with simple, declarative sentence structures.

3. *Create more items* than you will need so that you can choose those that best measure the important content and behaviors.

4. *Group the items* by same format (true-false, multiple-choice, essay) or by content or topical area.

5. *Review the items* after a "cooling off" period so that you can critique each item for its relevance to what you have actually taught (Thorndike, 1997; Worthen, et al., 1993).

6. *Prepare test directions carefully* to provide students with

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.

information about the test's purpose, time allowed, procedures for responding to items and recording answers, and scoring/grading criteria (Gronlund & Linn, 1990; McMillan, 2001).

Advantages and Disadvantages of Commonly Used Types of Assessment Items

As noted in Table 1, each type of exam item has its advantages and disadvantages in terms of ease of design, implementation, and scoring, and in its ability to measure different aspects of students' knowledge or skills. The following sections of this paper highlight general guidelines for developing multiple-choice and essay items. Multiple-choice and essay items 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 items (McMillan, 2001; Worthen, et al., 1993).

General Guidelines for Developing Multiple-Choice Items

Multiple-choice items have a number of *advantages*. First, multiple-choice items 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 can assess higher-order thinking skills as shown in Example 1, in which students are required to generalize, analyze, and make inferences about data in a medical patient case.

Multiple-choice items 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 for the instructor 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 (Gronlund & Linn, 1990; McMillan, 2001).

Example 1: A Series of Multiple-Choice Items That Assess Higher Order Thinking:

Patient WC was admitted for 3rd 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.

Two days later, WC's labs showed:
WBC 18,000 cells/mm3; 75% neutrophils (20% band cells); 15% lymphocytes; 6% monocytes; 2% eosophils; and 2% basophils.

Which one of the following best describes WC's lab results?

a. Leukocytosis with left shift
b. Normal neutrophil count with left shift
c. High eosinophil count in response to allergic reactions
d. High lymphocyte count due to activation of adaptive immunity

(Jeong Park, U-M College of Pharmacy, personal communication, February 4, 2008)

Guidelines for developing multiple-choice items

There are nine primary guidelines for *developing* multiple-choice items (Gronlund & Linn, 1990; McMillan, 2001). Following these guidelines increases the validity and reliability of multiple-choice items that one might use for quizzes, homework assignments, and/or examinations.

The first four guidelines concern the item "stem," which poses the problem or question to which the choices refer.

1. Write the stem as a clearly described question, problem, or task.
2. Provide the information in the stem and keep the options as short as possible.
3. Include in the stem only the information needed to make the problem clear and specific.

The stem of the question should communicate the nature of the task to the students and present a clear problem or concept. The stem of the question should provide only information that is relevant to the problem or concept, and the options (distractors) should be succinct.

4. Avoid the use of negatives in the stem (use only when you are measuring whether the respondent knows the exception to a rule or can detect errors).

You can word most concepts in positive terms and thus avoid the possibility that students will overlook terms of "no, not, or least" and choose an incorrect option not because they lack the knowledge of the concept but because they have misread the stated question. Italicizing, capitalizing, using bold-face, or underlying the negative term makes it less likely to be overlooked.

The remaining five guidelines concern the choices from which students select their answer.

5. Have ONLY one correct answer.

Make certain that the item has one correct answer. Multiple-choice items usually have at least three incorrect options (distractors).

6. Write the correct response with no irrelevant clues.

A common mistake when designing multiple-choice questions is to write the correct option with more elaboration or detail, using more words, or using general terminology rather than technical terminology.

7. Write the distractors to be plausible yet clearly wrong.

An important, and sometimes difficult to achieve, aspect of multiple-choice items is ensuring that the incorrect choices (distractors) appear to be possibly correct. Distractors are best created using common errors or misunderstandings about the concept being assessed, and making them homogeneous in content and parallel in form and grammar.

8. Avoid using "all of the above," "none of the above," or other special distractors (use only when an answer can be classified as unequivocally correct or incorrect).

All of the above and none of the above are often added as answer options to multiple-choice items. This technique requires the student to read all of the options and might increase the difficulty of the items, but too often the use of these phrases is inappropriate. *None of the above* should be restricted to items of factual knowledge with absolute standards of correctness. It is inappropriate for questions where students are asked to select "the best" answer. *All of the above* is awkward in that many students will choose it if they can identify at least one of the other options as correct and therefore assume all of the choices are correct – thereby obtaining a correct answer based on partial knowledge of the concept/content (Gronlund & Linn, 1990).

9. Use each alternative as the correct answer about the same number of times.

Check to see whether option "a" is correct about the same number of times as option "b" or "c" or "d" across the instrument. It can be surprising to find that one has created an exam in which the choice "a" is correct 90% of the time. Students quickly find such patterns and increase their chances of "correct guessing" by selecting that answer option by default.

Checklist for Writing Multiple-Choice Items

- ✓ Is the stem stated as clearly, directly, and simply as possible?
- ✓ Is the problem self-contained in the stem?
- ✓ Is the stem stated positively?
- ✓ Is there only one correct answer?
- ✓ Are all the alternatives parallel with respect to grammatical structure, length, and complexity?
- ✓ Are irrelevant clues avoided?
- ✓ Are the options short?
- ✓ Are complex options avoided?
- ✓ Are options placed in logical order?
- ✓ Are the distractors plausible to students who do not know the correct answer?
- ✓ Are correct answers spread equally among all the choices?

(McMillan, 2001, p. 150)

General Guidelines for Developing and Scoring Essay Items

Essays 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). Table 2 provides examples of essay question stems for assessing a variety of reasoning skills.

In addition to measuring complex thinking and reasoning, *advantages* of essays include the potential for motivating better study habits and providing the students

Table 2: Sample Essay Item Stems for Assessing Reasoning Skills

Skill	Stem
Comparing	Describe the similarities and differences between... Compare the following two methods for...
Relating Cause and Effect	What are the major causes of... What would be the mostly likely effects of...
Justifying	Which of the following alternatives do 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 for the following data. State a set of principles that can explain the following events.
Inferring	In light of the information presented, what is most likely to happen when... How would person X be likely to react to the following issue?
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/to... Describe what would happen if...
Applying	Using the principles of...as a guide, describe how you would solve the following problem. Describe a situation that illustrates the principle of...
Analyzing	Describe the reasoning errors in the following paragraph. List and describe the main characteristics of...
Synthesizing	Describe a plan for providing that... Write a well-organized report that shows...
Evaluating	Describe the strengths and weaknesses of... Using the given criteria, write an evaluation of...

Adapted from Figure 7.11 of McMillan, 2001, p.186.

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. It should, however, be noted that creating a high quality essay item takes a significant amount of skill and time.

The major *disadvantages* of essays include the amount of time faculty 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).

Types of essay items

We can distinguish between two types of essay questions, *restricted-response* and *extended-response*. A restricted-response essay focuses on assessing basic knowledge and understanding, and generally requires a relatively brief written response. Restricted-response essay questions assess topics of limited scope, and the nature of the question confines the form of the written response (see Example 2).

Example 2: Restricted-Response Items

State two hypotheses about why birds migrate. Summarize the evidence supporting each hypothesis.

(Worthen, et al., 1993, p. 277)

Identify and explain the significance of the following terms. For each term, be sure to define what it means, explain why it is important in American politics, and link it to other related concepts that we have covered in class. 1) Conformity costs 2) Voting Age Population (VAP)

(Charles Shipan, UM Department of Political Science, personal communication, February 4, 2008)

Extended-response essay items allow students to construct a variety of strategies, processes, interpretations, and explanations for a given question, and to provide any information they consider relevant (see Example 3). The flexibility of an extended-response item can make it less efficient for measuring specific learning outcomes than a restricted-response item, but it allows for greater opportunity to assess students' organization, integration, and evaluation abilities.

Example 3: Extended-Response Items

Compare and contrast the social conditions, prevailing political thought, and economic conditions in the U.S. North and South just prior to the outbreak of the Civil War; defend the issue that you believe was the most significant catalyst to the war.

(Worthen, et al., 1993, p. 276)

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?

(Charles Shipan, UM Department of Political Science, personal communication, February 4, 2008)

Guidelines for developing essay items

There are six primary guidelines for developing essay items (Gronlund & Linn, 1990; McMillan, 2001; Worthen, et al., 1993).

1. Restrict the use of essay questions to educational outcomes that are difficult to measure using other formats.
2. Construct the item to elicit skills and knowledge in the educational outcomes.
3. Write the item so that students clearly understand the specific task.

Other assessment formats are better for measuring recall knowledge (e.g., true-false, fill-in-the-blank, multiple-choice); the essay is able to measure deep understanding and mastery of complex information. When constructing essay items, start by identifying the specific skills and knowledge that will be assessed. As noted earlier, Table 2 provides examples of essay item question stems for assessing a variety of reasoning skills. Building from these simple stem formats, you can create a restricted-response

or extended-response item to assess one or more skills or content topics.

Once you have identified the specific skills and knowledge, you should word the question clearly and concisely so that it communicates to the students the specific task(s) you expected them to complete (e.g., state, formulate, evaluate, use the principle of, create a plan for, etc.). If the language is ambiguous or students feel they are guessing at "what the instructor wants me to do," the ability of the item to measure the intended skill or knowledge decreases.

4. Indicate the amount of time and effort students should spend on each essay item.

In essay items, especially when used in multiples and/or combined with other item formats, you should provide students with a general time limit or time estimate to help them structure their responses. Providing estimates of length of written responses to each item can also help students manage their time, providing cues about the depth and breadth of information that is required to complete the item. In restricted-response items a few paragraphs are usually sufficient to complete a task focusing on a single educational outcome.

5. Avoid giving students options as to which essay questions they will answer.

A common structure in many exams is to provide students with a choice of essay items to complete (e.g., "choose two out of the three essay questions to complete…"). Instructors, and many students, often view essay choice as a way to increase the flexibility and fairness of the exam by allowing students to focus on those items for which they feel most prepared. However, the choice actually *decreases the validity and reliability* of the instrument because *each student is essentially taking a different test*.

Creating parallel essay items (from which students choose a subset) that test the same educational objectives (skills, knowledge) is very difficult, and unless students are answering the same questions that measure the same outcomes, scoring the essay items and the inferences made about student ability are less valid. While allowing students a choice gives them the perception that they have the opportunity to do their best work, you must also recognize that choice entails *difficulty in drawing consistent and valid conclusions* about student answers and performance.

6. Consider using several narrowly focused items rather than one broad item.

For many educational objectives aimed at higher order reasoning skills, creating a series of essay items that elicit different aspects students' skills and knowledge can be more efficient than attempting to create one question to capture multiple objectives. By using multiple essay items (which *all* students complete), you can capture a variety of skills and knowledge while also covering a greater breadth of course content.

Guidelines for scoring essay items

There are five primary guidelines for scoring essay items (Gronlund & Linn, 1990; McMillan, 2001; Wiggins, 1998; Worthen, et al., 1993; *Writing and grading essay questions*, 1990).

1. Outline what constitutes an expected answer (criteria for knowledge and skills).

Developing the criteria for what you expect students to know and be able to demonstrate in advance of giving the exam allows you to refine the wording of the essay questions so that they best elicit the content, style, and format you desire.

Identifying the criteria in advance also decreases the likelihood that you will be influenced by the initial answers you read when you begin to grade the exam. It is easy to become distracted by the first few answers one reads for a given essay item and allow the skills and knowledge (or lack thereof) demonstrated in these few students' answers to set the scoring criteria, rather than relying on sound criteria based on the educational outcomes of the course curricula. Using predetermined criteria increases the fairness, accuracy, and consistency of the scoring process.

2. Select an appropriate scoring method based on the criteria.

Generally we describe scoring methods for written communication as either *trait analytic*, where each identified criterion is assigned separate points, or *holistic*, where a single score is generated resulting in an overall judgment of the quality. Restricted-response essays are often graded using the trait analytic method and extended-response essays, the holistic method, but either method can be used with either item type. To implement your approach, you can develop a *rubric* or scoring key that specifies the educational criteria for scoring (as described in the previous

guideline) and the amount of credit or points to be assigned for each criterion.

The validity of a scoring rubric is affected by two factors: how clearly the descriptions in the rubric's categories discriminate between levels of student performance; and the degree of congruence between the criteria and the knowledge and skills imbedded in the instruction. In other words, do the criteria outlined in the rubric genuinely reflect the tasks that are important for answering the question (use of examples, rationale for inferences, etc.)? Or do they instead describe general behaviors useful when answering any essay question (engaging tone, interesting format, etc.) but not critical to the specific tasks at hand? Example 4 shows a rubric for grading critical papers that clearly and succinctly describes expectations for conceptual structure, rhetorical structure, thesis statement, evidence and analysis, paragraph/section structure, and sentence mechanics for five levels of grades (A-D, F).

Finally, effective rubrics should avoid the use of comparative language such as "better than, more than, worse than, less than…" in discriminating between the levels of performance. Rather, the rubric should describe specifically those characteristics that define and delineate that level of performance, as is shown in Example 5, a rubric for four levels of performance on critical thinking.

Example 5: Levels of Performance for Critical Thinking

4 = Exemplary: Clearly defines the issue or problem; accurately identifies the core issues; appreciates depth and breadth of problem. Identifies and evaluates relevant significant points of view.
3 = Satisfactory: Defines the issue; identifies the core issues, but does not fully explore their depth and breadth. Identifies and evaluates relevant general points of view.
2 = Below Satisfactory: Defines the issue, but poorly (superficially, narrowly); overlooks some core issues. Identifies other points of view but focuses on insignificant points of view.
1 = Unsatisfactory: Fails to clearly define the issue or problem; does not recognize the core issues. Ignores or superficially evaluates alternate points of view.

(Adapted from *Critical Thinking Rubric*, 2008)

Trait analytic scoring provides students with specific feedback about each criterion and a summed score out of the possible total points, but can be very time consuming for the instructor to implement. Restricting the educational criteria to only those most essential for an individual essay question, perhaps 4-6 criteria, and clearly outlining how points will be awarded (or deducted) streamlines the process (McMillan, 2001).

Concerns about fairness, accuracy, and consistency of the essay scoring process increase when more than one grader has responsibility for scoring essay items. Such a situation often occurs in large courses where the faculty member and multiple graduate student instructors share the responsibility. Developing criteria and scoring rubrics together and defining how the scoring rubrics will be implemented helps to establish common understandings about expectations for students' work. It is also helpful to have each member of the group grade a small subset of essay responses and then meet together to compare and calibrate their scoring to ensure consistency before completing the scoring for the entire set of exams (*Writing and grading essay questions*, 1990).

3. Clarify the role of writing mechanics and other factors independent of the educational outcomes being measured.

For example, you can outline for students how various elements of written communication (grammar, spelling, punctuation, organization and flow, use of vocabulary/ terminology, use of scientific notation or formulas) figure into the scoring criteria. You should also decide whether you will decrease scores or ignore the inclusion of information irrelevant to the question.

4. Use a systematic process for scoring each essay item.
5. Create anonymity for students' responses while scoring.

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. In other words, when faced with a set of exams to score, read and score all of essay question #1, then shuffle the set of exams to randomly reorder them, and then read and score all of essay question #2. Grading the same essay question for all students creates a more uniform standard of scoring as it is easier to remember the criteria for scoring each answer. Shuffling decreases the effect of grader fatigue on the accuracy of the scoring, as exams that might otherwise be scored at the end of the pile (e.g., students with last names at the end of the alphabet or higher I.D. numbers)

Example 4: Grading Rubric for Critical Papers

Letter Grades	Conceptual Structure	Rhetorical Structure	Thesis	Evidence and Analysis	Paragraph/Section Structure	Sentence Mechanics
A	cogent analysis, shows command of interpretive and conceptual tasks required by assignment and course materials: ideas original, often insightful, going beyond ideas discussed in lecture and class	contains a convincing argument with a compelling purpose; highly responsive to the demands of the specific writing situation; sophisticated use of conventions of academic discipline and genre; anticipates the reader's need for information, explanation, and context	essay controlled by clear, precise, well-defined thesis; is sophisticated in both statement and insight	well-chosen examples; uses persuasive reasoning to develop and support thesis consistently; uses specific quotations, aesthetic details, or citations of scholarly sources effectively; logical connections between ideas are evident	well-constructed paragraphs; appropriate, clear, and smooth transitions; apt arrangement of organizational elements	uses sophisticated sentences effectively; usually chooses words aptly; observes professional conventions of written English and manuscript format; makes very few minor or technical errors
B	shows a good understanding of the texts, ideas and methods of the assignment; goes beyond the obvious; may have one minor factual or conceptual inconsistency	addresses audience with a thoughtful argument with a clear purpose; responds directly to the demands of a specific writing situation; competent use of the conventions of academic discipline and genre; addresses the reader's need for information, explanation, context	clear, specific, arguable thesis central to the essay; may have left minor terms undefined	pursues explanation and proof of thesis consistently; develops a main argument with explicit major points, appropriate textual evidence, and supporting detail	distinct units of thought in paragraphs controlled by specific, detailed, and arguable topic sentences; clear transitions between developed, cohering, and logically arranged paragraphs	a few mechanical difficulties or stylistic problems; may make occasional problematic word choices or syntax errors; a few spelling or punctuation errors; usually presents quotations effectively, using appropriate format
C	shows an understanding of the basic ideas and information involved in the assignment; may have some factual, interpretive, or conceptual errors	presents adequate response to the essay prompt; pays attention to the basic elements of the writing situation; shows sufficient competence in the conventions of academic discipline and genre; signals the importance of the reader's need for information, explanation, and context	general thesis or controlling idea; may not define several central terms	only partially develops the argument; shallow analysis; some ideas and generalizations undeveloped or unsupported; makes limited use of textual evidence; fails to integrate quotations appropriately; warrants missing	some awkward transitions; some brief, weakly unified or undeveloped paragraphs; arrangement may not appear entirely natural; contains extraneous information	frequent wordiness; unclear or awkward sentences; imprecise use of words or over-reliance on passive voice; contains rudimentary grammatical errors; makes effort to present quotations accurately
D	shows inadequate command of materials or has significant factual and conceptual errors; confuses some significant ideas	shows serious weaknesses in addressing an audience; unresponsive to the specific writing situation; poor articulation of purpose in academic writing; often states the obvious or the inappropriate	thesis vague or not central to argument; central terms not defined	frequently only narrates; digresses from one topic to another without developing ideas or terms; makes insufficient or awkward use of textual evidence; relies on too few or the wrong type of sources	simplistic, tends to narrate or merely summarize; wanders from one topic to another; illogical arrangement of ideas	some major grammatical or proofreading errors (subject-verb agreement, sentence fragments, word form errors, etc.); repeated inexact word choices; incorrect quotation or citation format
F	lacks critical understanding of lectures, readings, discussions, or assignments	shows severe difficulties communicating through academic writing	no discernible thesis	little or no development; may list disjointed facts or misinformation; uses no quotations or fails to cite sources or plagiarizes	no transitions; incoherent paragraphs; suggests poor planning or no serious revision	numerous major and minor grammatical errors and stylistic problems; does not meet Standard Written English requirement

(T. McElroy and F. Whiting, University of Alabama, personal communication, February 15, 2008)

are shuffled throughout the scoring process and scored randomly among the others.

The shuffling also creates a random order in which to grade each essay item, making it less likely that you will identify a pattern in an individual student's answers or base your score on previous impressions of that student. You might also choose to completely cover students' names/I.D. numbers so that you cannot identify the student until after you have graded the entire exam.

Checklist for Writing Essay Items

- ✓ Is the targeted reasoning skill measured?
- ✓ Is the task clearly specified?
- ✓ Is there enough time to answer the questions?
- ✓ Are choices among several questions avoided?

Checklist for Scoring Essays

- ✓ Is the answer outlined before testing students?
- ✓ Is the scoring method–holistic or trait analytic–appropriate?
- ✓ Is the role of writing mechanics clarified?
- ✓ Are items scored one at a time?
- ✓ Is the order in which the tests are graded changed?
- ✓ Is the identity of the student anonymous?

(McMillan, 2001, p. 185, 188)

You can also use these guidelines for scoring essay items to create scoring processes and rubrics for students' papers, oral presentations, course projects/products/artifacts, and websites/technological tools.

Normative and Criterion Grading Systems

Fair and accurate grading depends on the development of effective and well-designed exams and assignments as described above. Just as important to the process is the approach you take to assigning grades. In general, we can divide these approaches to grading into two main categories: *norm-referenced* and *criterion-referenced* systems. Whatever system you use needs to be fair, distinguish clearly between different levels of achievement, and enable students to track their progress in the course (*Grading systems*, 1991).

Norm-referenced grading

When using a norm-referenced approach, faculty compare a student's performance to the norm of the class as a whole. A normative system assumes that skills and knowledge are distributed throughout the population (norm group). Assessment is designed to maximize the difference in student performance, for example by avoiding exam questions that almost all students will answer correctly (Brown, 1983; *Grading systems*, 1991; Gronlund & Linn, 1990; Isaac & Michael, 1990; Svinicki, 1999a; Svinicki, 1999b). Advocates of norm-referenced grading argue it has several benefits. Norm-referenced grading allows for flexibility in assigning grades to cohorts of students that are unusually high- or low-achieving, thus allowing distinctions to be made. It also allows faculty to correct for exams that turn out to be much more or less difficult than anticipated. Norm referenced grading can potentially combat grade inflation by restricting the number of students receiving the highest grades. Finally, this approach can be valuable because it identifies students who stand out in their cohort, which can be helpful for faculty writing letters of recommendation.

The most common version of this approach is grading on the curve, in which grades on an exam are placed on a standard bell curve, in which, for example, 10% receive "A" or "F," 25% receive "B" or "D," and 35% receive the average grade of "C." However, a strict bell curve demonstrates the major drawbacks of norm-referenced grading, especially in relationship to student motivation and cooperation. Students realize in normative grading that some students *must* receive low/failing grades. It can be difficult for an individual student to gauge his/her performance during the semester and develop strategies for remediation because final grades are dependent upon how all others perform in total and not just on that individual student's skills and abilities (Brown, 1983; *Grading systems*, 1991; Gronlund & Linn, 1990; Isaac & Michael, 1990; Svinicki, 1999a; Svinicki, 1999b). Because students are competing for grades, faculty may also find it very difficult to implement any type of cooperative learning or peer support. In practice, faculty rarely use a strict bell curve in which 35% would receive a D or F. However even a more forgiving curve has the potential to dampen motivation and cooperation in a classroom.

The consequences, especially for science majors, can be significant, causing students to question whether they should persist in their major. Studies on persistence in the sciences have found that competitive grading designed to sort students – especially in introductory, gateway courses – can have a significant impact on students' decisions to leave the sciences. The effect is particularly pronounced for underrepresented students such as women and minorities in the sciences (Seymour & Hewitt, 1997).

Criterion-referenced grading

Criterion-referenced grading focuses on the *absolute performance* against predetermined criteria. Criterion systems assume that grades should measure "how much" a student has mastered a specified body of knowledge or set of skills. Assessment items are created to ensure content coverage, irrespective of whether they are easy or difficult (Brown, 1983; *Grading systems*, 1991; Gronlund & Linn, 1990; Isaac & Michael, 1990; Svinicki, 1999a; Svinicki, 1999b). This approach to grading rewards students for their effort and conveys that anyone in the class can achieve excellent results if they meet the standards. As a result, it is well suited to collaboration rather than competition. It can also support mastery learning (Svinicki, 1998), especially if students understand the criteria in advance so that they can direct their studying accordingly.

Criterion-referenced grading does have drawbacks as well. Students may perceive as arbitrary the system by which cut scores are determined in criterion systems (e.g., an A is 90% and above, thus 89% is a B) and may aggressively pursue faculty for "extra credit" assignments or lobby for additional points. Criterion systems also require instructors to consider the relative difficulty of each assessment and weight them accordingly (i.e., more difficult or comprehensive assessments should be given more weight). In a poorly designed criterion system, a student who performs well on simple, less rigorous assessments (quizzes, homework) may achieve a higher grade than a student who performed well on more difficult assessments (midterm/final exams) if the simple assignments are weighted more heavily in the final grade calculation (*Grading systems*, 1991). Criterion systems are sometimes seen as contributors to grade inflation, as it is possible for all students to achieve grades of A if they meet the criteria for proficiency.

In practice, most faculty will use some combination of systems to determine student grades. Faculty using a norm-referenced approach may look at the quality of student work when setting the cutoff points on a curve. When using a criterion-referenced system, faculty may look at the numbers of students receiving very high or very low scores to help them determine how they will set the actual grades. As you make choices about grading systems it might be worth considering the assumptions that underlie your own view of the purpose of grading. As McKeachie and Svinicki (2006) frame the issue, "Is the purpose of the grade to identify the 'best' students in a group (norm referencing), or is it to indicate what each student has achieved (criterion referencing). Both are legitimate positions and can be and are argued for vociferously" (p. 132).

Conclusion

Regardless of the type of assessment instrument developed to measure students' knowledge and skill in college-level courses, it is important to remember to explain the purpose of assessment to ourselves and our students, understand the importance of using valid and reliable instruments, and carefully consider judgments made about data from those assessment instruments. By following guidelines for designing, implementing, and scoring exams, instructors can enhance the validity, reliability, and utility of the information they collect about students' learning. High quality assessment provides instructors and students opportunities for more meaningful and relevant teaching and learning in the classroom.

References

Brown, F. G. (1983). *Principles of educational and psychological testings* (3rd ed.). New York: Holt, Rinehart and Winston.

Cashin, W. E. (1987). *Improving essay tests.* Idea Paper, No. 17. Manhattan, KS: Center for Faculty Evaluation and Development, Kansas State University.

Critical thinking rubric. (2008). Dobson, NC: Surry Community College.

Grading systems. (1991, April). For Your Consideration, No. 10. Chapel Hill, NC: Center for Teaching and Learning, University of North Carolina at Chapel Hill.

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.

Isaac, S., & Michael, W. B. (1990). *Handbook in research and evaluation.* San Diego, CA: EdITS Publishers.

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.

Seymour, E., & Hewitt, N. M. (1997). *Talking about leaving: Why undergraduates leave the sciences.* Boulder, CO: Westview Press.

Svinicki, M. D. (1998). Helping students understand grades. *College Teaching, 46*(3), 101-105.

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.

Svinicki, M. D. (1999b). Some pertinent questions about grading. *In Teachers and students: A sourcebook for UT-Austin faculty* (pp. 1-2). 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.

The *CRLT Occasional Papers* series is published on a variable schedule by the Center for Research on Learning and Teaching at the University of Michigan. Information about extra copies or back issues can be obtained by writing to: Publications, CRLT, 1071 Palmer Commons, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218.

CRLT Occasional Paper No. 24

Center for Research on Learning and Teaching

The University of Michigan
1071 Palmer Commons
100 Washtenaw Avenue
Ann Arbor, MI 48109-2218