CENTER FOR FACULTY EXCELLENCE PRESENTS

for your consideration...

SUGGESTIONS AND REFLECTIONS ON TEACHING AND LEARNING

March 2012

Writing and Grading Essay Questions

"Essay grades ought to reflect not just the quantity of knowledge accurately reproduced but also the quality of reasoning, logic or insight the student generates in the time allowed"

(Brown, 2010)

A hundred years ago, all college course exams were essay exams. The multiple-choice exam had not yet been invented, and students were expected to demonstrate their learning by producing detailed answers to question prompts in clear prose. Today many teachers still consider essay questions the preferred method of assessment. Arguments in favor of the essay exam include:

- Essays require both more effort and deeper understanding on the part of the student than do other types of questions. For example, students typically cannot produce an adequate answer to an essay question simply by recognizing or recalling the correct information, as they often can with multiple-choice or short-answer questions.
- Essay questions can better assess the complexity of students' thought processes, as well as their ability to think critically and solve problems in a particular domain of knowledge, than can multiple-choice or fill-in-the-blank items.
- Essays require students to express their thoughts in grammatical, well-crafted sentences and paragraphs—a skill that all educated persons should be expected to master.

Still, the essay form has its limitations. There are four major arguments against using essay questions to assess student learning:

1. ***Essays take much longer than other types of assessment items for students to write and for teachers to grade.*** Consequently, for a fixed amount of examination time or grading effort, fewer items can be administered, which weakens the validity of the exam. Specifically, an exam that does not adequately sample the entire domain of knowledge being assessed (i.e., what the student is supposed to have learned) may not yield an accurate indication of how well the student has mastered the course content. One way to achieve adequate domain coverage with essay exams is to administer exams that require many hours or days to complete (e.g., components of the Bar examination that assesses legal knowledge), but this is not a realistic option for most college instructors.

2. ***Grading of essay exams can be influenced by extraneous factors.*** These factors include such things as handwriting legibility and ink color (Klein & Taub, 2005; Joseph, 2005; Greifeneder et al., 2010). In addition, if the exams are not graded anonymously, teacher expectations regarding the quality of a student's work, stereotypes associated with the student's sex or ethnicity, or the grader's personal feelings toward the student may influence essay scoring (Chase, 1986; Hughes, Keeling, & Tuck,1983). Contrast or order effects may also play a role; essays preceded in the grading queue by poor quality papers tend to receive higher scores than do the same essays when preceded by high quality papers (Spear, 1997). Because these factors have no systematic relationship to the quality of the ideas expressed in a student's essay, their influence weakens the validity of the assigned score in that the score assigned to the essay does not accurately indicate the degree of subject mastery the student attained.

3. ***Essay grades are unreliable.*** Due in part to the influence of extraneous factors, there tends to be relatively poor agreement among graders in the score assigned to an essay. Moreover, even individual graders are often inconsistent in the scores they assign to an essay on two different readings. Specifically, in empirical studies of reliability in essay scoring, both consensus (i.e.,

exact numerical agreement between scores) and correlation (i.e., degree to which the essays in a set are ranked similarly by different graders) have been found to be low (Brown, 2010). The fact that different graders (or even the same grader in the case of multiple readings) often cannot agree on what score a given essay should receive undermines confidence that a student's score on an essay exam accurately reflects the student's mastery of the material being tested.

4. ***Essay exam performance conflates course subject matter knowledge with writing skill.*** Some authors argue that essay exams are a poor method of assessing subject matter knowledge because the student's performance—and, consequently, her exam score—depend not only on her knowledge of the subject matter being tested but also on her written communication skills.

Given these limitations, if there existed a satisfactory alternative way to assess the highest levels of understanding and reasoning in a domain of knowledge, we might well dispense with the essay exam altogether. At present, however, "objective" forms of assessment such as multiple-choice tests are at best a complement to essay exams and not an adequate substitute for them. Your task as an instructor, then, is to devise essay question prompts and grading procedures that will minimize the threats to validity and reliability described above. Below we offer some suggestions to help you avoid some of the more common pitfalls instructors encounter in creating and grading essay questions.

1. ***Determine whether an essay question is the most appropriate format for the type of learning you want to assess.*** Essay questions should be used when you want to assess students' ability to think critically and organize their thoughts, or to demonstrate their understanding by taking the factual information they have learned and applying it in some way. For example, an essay question might ask a student to critique an argument, interpret a text, justify a position on some issue, explain the causes of some phenomenon, or predict the effect of an intervention (see Reiner et al., 2003, for an extensive list of the different kinds of tasks you might ask students to perform on an essay exam).

2. ***Administer enough different question items to adequately sample the domain of knowledge covered by the exam.*** In order for a student's exam score to serve as a valid indicator of how well she has mastered the material covered by the exam, there must be a sufficient number of question items, addressing a sufficiently diverse set of topics, to adequately represent the full range of subject matter the students are expected to have learned. Single-item essay tests rarely meet this criterion unless they are broken down into a number of sub-components, in effect becoming a set of short essays. In many cases it is preferable to use a number of short essay questions to insure that the material has been sampled adequately.

3. ***Avoid ambiguous prompts; state the question clearly and precisely and make clear what information the answer should contain.*** Do not assume that your students will interpret a vaguely-worded question in the way you intend; it is better to err on the side of providing too much detailed guidance in the prompt than too little. In addition to specifying the information that an essay answer should contain, the prompt can help students allocate their time and effort appropriately by indicating how much time they should spend on each part of the answer and/or how many points each part of the answer is worth, as does the essay prompt from a Physical Anthropology exam shown in Figure 1. Note that the three parts of this prompt ask students to demonstrate progressively more complex and sophisticated forms of understanding and mastery. Part I requires primarily recall and comprehension, Part II requires application and analysis, and Part III requires synthesis and evaluation.

Figure 1

Lectures covering Piltdown Man, Gradualism, Punctuated Equilibrium, and Catastrophism were given sequentially to illustrate the interplay of theory and fact in the formulation of an anthropological account of the evolution of humankind. Write a three-part essay addressing the following questions:

I. Name the major proponents of the above underlined concepts and briefly describe the significance of these people for the history of a science of evolution. (10 minutes, 10 points)

II. Select any two of the four concepts above and explain how they illustrate the relationship between fact and theory. (10 minutes, 10 points)

III. In your opinion, are new discoveries or theories really new or are they just repetitions of past ideas that have fallen out of favor? Your answer to part III must draw upon the four concepts underlined above and be consistent with what you have already written in parts I and II. (20 minutes, 20 points)

4. If the purpose of the exam is to assess students' content knowledge rather than writing skill per se, ***provide a detailed prompt that mitigates the degree to which a student's writing skill interferes with her ability to demonstrate knowledge of the course subject matter.*** Brown (2010) suggested two approaches you might take. First, you might "require all essay writers to use the same organizational patterns so that structural characteristics will not be used by the [grader] as a proxy for knowledge and understanding in the content area" (p. 283). For example, Figure 2

shows an essay prompt administered to students in a course that covered the theories and methods of intelligence measurement. The instructor wanted to assess how well the students could use what they had learned in the course to analyze the theoretical model underlying a sample set of questions from a particular intelligence test (Brown, 2010). According to the instructor, "By following the two main questions and their subparts, the students could create a written response that clearly focused on the essential element of interest... without having to worry that their responses would be scored according to their essay organization skills" (p. 283).

Figure 2

Structured essay prompt, from Brown (2008, p. 53)

1. What are the intelligence factors assessed by these items?
    a. Identify the mental ability that each item type is testing.
    b. Explain why the item types group into the factors you have chosen.
    c. What labels from Carroll's taxonomy best describe your factors?
    d. What theorist or theory is most associated with the factor pattern you have chosen?
    e. Why are these factors important measures of intelligence?
2. Explain what kind of relationship you would expect there to be between the factors you have identified.
    a. What kind of correlation and/or factor pattern will there be?
    b. What kind of hierarchy, if any, will there be?
    c. What does this relationship pattern say about the nature of intelligence as measured by the test?

A second approach is to provide in the prompt all the sequencing or transition phrases the student is expected to use in answering the question. For example, Figure 3 shows a prompt used by a professor of Education. The prompt is designed to guide students in writing an essay that responds to the statement "Self concept and academic achievement are not related."

Figure 3

Structured essay prompt, from Brown (2008, p. 54)

- The evidence on this topic generally says . . .
- Four contrasting findings in the literature on this topic are . . .
- How do these studies aid in addressing the topic?
- Why is it more beneficial to assess how self-concept relates to learning?
- How does self-concept influence learning and learning influence self-esteem?
- What strategies do students use to maintain their status quo sense of self-esteem?
- Note some teaching procedures you, as a teacher, could use to redress these strategies.

5. ***Do not allow students to choose which questions to answer.*** There is little point in taking the trouble to devise a set of essay prompts that adequately sample the subject matter covered by the course, only to then let your students select a subset of the items to answer. They understandably will choose those items they feel most capable of answering, and the students' exam scores therefore will not accurately indicate the students' mastery of all the subject matter covered by the test.

6. ***Minimize influence of extraneous factors.*** To reduce the effect of teacher expectancy, instruct the students to write their PID number (or some other identifier) on their exam rather than their name. If you require your students to write their actual name on the cover of the exam booklet (e.g., to sign the honor pledge), fold over the front page of all the exams before you begin grading. To reduce the influence of handwriting legibility, ink color, or other surface characteristics, consider having students compose their essays on their laptops. Software programs now exist (e.g., "Electronic Blue Book") that effectively turn students' laptops into typewriters by restricting access during the exam to the internet or to files and applications saved on their computers. This approach enables students with superior typing skills to write longer essays in a fixed period of time than those without such skills, so you may need to allow slow typists extra time. To reduce contrast and order effects, it is advisable to grade each exam at least twice, perhaps employing multiple graders, and to randomize the order in which exams are graded each time.

7. ***Adopt practices to improve the reliability of grades.*** Before beginning to grade, write a model answer to each exam item. The model answer contains all the information expected in a full-credit answer. A more systematic and analytic method of creating model answers is to prepare a grading rubric (see FYC #4). Quickly skim several essays before beginning the formal process of grading to determine whether or not the model answer needs to be modified. If, through some quirk in wording, students misinterpret your intent, or if your standards are unrealistically high (or low), you should alter the model answer in light of this information. This procedure is preferable to altering the grading scheme retrospectively, since grades tend to lose their meaning if the system is altered to compensate for poor testing practices.

    Grade each essay question separately rather than grading a student's entire test at once. A brilliant performance on the first question may overshadow weaker answers later on (or vice-versa), and it is easier for you to keep in mind one model answer at a time. Shuffling

the papers after grading each question will help compensate for the tendency to give later papers lower scores as you tire, and will further help to reduce contrast and order effects.

Unless elements of grammar, syntax, spelling, and punctuation are being evaluated as part of the examination, try to overlook flaws in these elements of composition. If mastery of subject matter and not quality of written expression is the target of assessment, accuracy and completeness should be the only criteria against which the answers are judged.

Final Thoughts

The authority ascribed to essay questions derives in part from the weight of tradition—we've used them for so long that it is difficult to imagine any other kind of assessment. However, as discussed above, there are many ways in which an essay exam can fall short as an assessment instrument, and it requires considerable thought and effort to create an exam that achieves satisfactory levels of validity and reliability. Indeed, a well-designed essay exam is a work of scholarship in its own right. Therefore, invest the time and energy needed to create essay prompts and grading procedures that you can feel as proud of as you would an article submitted for publication. Your students deserve no less.

Checklist for Writing and Grading Essay Exams

- Are essays the appropriate means to test the material you have covered?
- Have you been using essay-type questions throughout the semester as means of generating discussion in class?
- If there is a choice of questions, are they truly equivalent? Would it be better to have several short essays?
- What are your specific grading criteria? Have you made these criteria clear in the instructions?
- Have you provided for anonymous grading?
- Do you have a model answer or rubric against which you can judge student responses?
- If there are several TAs, are the grading criteria clear to all involved?
- Do you intend to discuss the exam when you return it?

Bibliograhy

Brown, G. T. L. (2008). *Conceptions of assessment: Understanding what assessment means to teachers and students*. New York, NY: Nova Science.

Brown, G. T. L. (2010). The validity of examination essays in higher education: Issues and responses. *Higher Education Quarterly*, 64(3), 276-291.

Chase, C. I. (1986). Essay test scoring: Interaction of relevant variables. *Journal of Educational Measurement*, 23(1), 33-41.

Davis, B. G. (2009). *Tools for teaching (2nd ed.)*. San Francisco, CA: John Wiley & Sons.

Greifeneder, R., Alt, A., Bottenberg, K., Seele, T., Zelt, S., & Wagener, D. (2010). On writing legibly. *Social Psychological and Personality Science*, 1(3), 230.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1983). Effects of achievement expectations and handwriting quality on scoring essays. *Journal of Educational Measurement*, 20(1), 65-70.

Klein, J., & Taub, D. (2005). The effect of variations in handwriting and print on evaluation of student essays. *Assessing Writing*, 10(2), 134-148.

Reiner, C. M., Bothell, T. W., & Sudweeks, R. R. (Eds.). (2003). *Preparing effective essay questions. Stillwater*, OK: New Forums Press.

Spear, M. (1997). The influence of contrast effects upon teacher's marks. Educational Research, 39(2), 229. Retrieved from https://auth.lib.unc.edu/ezproxy_auth.php?url=http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=9708123423&site=ehost-live&scope=site

UNC CENTER FOR FACULTY EXCELLENCE

316 WILSON LIBRARY; CAMPUS BOX 3470
CHAPEL HILL, NC 27599-3470

919.966.1289; cfe@unc.edu
http://cfe.unc.edu