Running head: ASSESSING LEARNING ONLINE

Assessing Learning Online: The Top Ten List

Beverly M. Klecker

Morehead State University

Morehead, KY, USA

b.klecker@morehead-st.edu

Paper presented at the annual meeting of

The Society for Information Technology

and Teacher Education (SITE)

Phoenix, AZ

March 5, 2005

Abstract

The purpose of instruction, whether face-to-face or online, is to facilitate student achievement of intended learning objectives. Traditionally, the major function of classroom assessment in undergraduate and graduate university courses has been to measure the individual student's learning in order to provide feedback to the student and to spread student scores (norm-referenced grading) to assign grades. Online assessments can measure the student's achievement of intended learning objectives if, and only if, great diligence is used in their construction. This paper will address the top ten issues in designing online classroom assessment: (1) measuring the objectives, (2) cognitive levels, (3) ethical considerations, (4) formative/summative, (5) criterion referenced /norm-referenced, (6) validity, (7) reliability, (8) diversity, (9) variety, (10) providing feedback.

Assessing Learning Online: The Top Ten List

One: Measuring the Objectives

Traditionally, the major function of classroom assessment in undergraduate and graduate university courses has been to measure the individual student's learning in order to provide feedback to the student and to spread student scores to assign grades (Sax, 1997). Too often, the instructor views as secondary the question of whether or not the intended objectives have been achieved (Brookhart, 2000). Before beginning instruction, the instructor should always determine the learning outcomes expected from the course. Ideally, these are the course objectives delineated in the course syllabus. Gronlund (2000) suggested that instructors consider the *level* of cognitive outcomes expected within each content outcome. Bloom, Englehart, Frost, Hill, & Krathwohl (1956) presented six cognitive levels in a their well-known hierarchical taxonomy. From lowest to highest they are (1) knowledge, (2) understanding, (3) application, (4) analysis, (5) synthesis, and (6) evaluation.

Two: Cognitive Levels

All instruction has some objectives aimed at Bloom's lower cognitive levels. Such objectives would include, for example, learning the new vocabulary of a discipline and understanding how to apply the vocabulary. Yet, even introductory courses should require students to analyze, synthesize, and evaluate new concepts while considering previous learning. In introductory courses with much new material and vocabulary, the first assessments in the course will included many lower-level knowledge and understanding level questions. Multiple-choice items are ideal for measuring the achievement of objectives at Bloom's three lower cognitive levels. The instructor may wish to assess the students' ability to apply this new

knowledge by using application-type questions containing material that has not been used as examples in instruction.

Requiring the student's composition of an essay, a poem, or a musical arrangement will provide a measure of synthesis. Asking a student to diagram the structure of sentences for an English course is an excellent example of analysis. Evaluating (determining the worth) a paragraph, a painting, an argument, or a portfolio selection is, according to Bloom's cognitive taxonomy, the highest cognitive level. Yet, if you've not planned to teach these "higher-order" thinking skills, and if you've not provided activities to teach the thinking skills, then you must *not* assess them! Remember, teach what you assess; assess what you teach. This is a very straight-forward principle that is ignored by many.

Three: Ethical Issues

Benson (2003, p. 70) suggested, "Two key benefits of online assessments are (1) the ability of every learner to respond to every question the instructor asks and (2) the ability of the instructor to provide immediate feedback to each learner. In a traditional course, when the instructor asks a question, the first student to answer is typically afforded the sole opportunity to provide an answer." However, how can we be sure that the responses that we receive are coming from our students? As an online instructor, I faced this early on when I used 20 multiple-choice items a week to focus the student's reading in the assigned chapter. Naively, I placed the correct response for the student's viewing when he or she received immediate scoring feedback on Blackboard. I then noticed that the Discussion Board was being used to relay the correct answers! A quick solution to this was to provide the answers as feedback on Monday morning--after all student responses had been received by the Sunday 11:55 p.m. deadline. I also require

all online students to take a tutorial on plagiarism and to submit a statement that they understand the definition of plagiarism and agree that all work submitted in their name is their work.

Constructing unique assignments (beyond term papers) that require unique responses will also help students resist plagiarism temptation. A new feature on Blackboard version 6.0 allows me to "scramble" the items of the multiple-choice tests. Changing items every semester works even better, but care must be taken to retain the variety of cognitive levels. Northcote (2002, p. 624) suggested, "... it is the driving force of the pedagogical beliefs of the users of such systems that will ultimately reflect the quality of online assessment. Schneider (2002) referred to online assessment as an ethical minefield. This is an apt metaphor and instructors must either take care to remove the "mines" before entering, or tread very gingerly indeed. Ricketts and Wilks (2002) presented results of research with a specific class that indicated that students may be disadvantaged by the introduction of online assessment, unless care is taken with the student-assessment interface. Online instructors should carry out their own action research studies to determine the assessment parameters (e.g., presentation mode, number of items, timing of assessment) that result in optimal performance.

Four: Formative and Summative

Formative and summative assessment are often distinguished in the field of evaluation by *time*--that is when evaluation occurs in instruction. Formative assessment (often the "midterm") is used to provide feedback to the students and instructors; summative assessment (the "final") is used to determine whether the student will pass or fail the course. Brookhart (2004, p. 45) succinctly summarized the difference between formative and summative assessment,

> Formative assessment means information gathered and reported
> for use in the development of knowledge and skills, and summative

> assessment means information gathered and reported for use in judging the outcome of that development. As the saying goes,
>
> 'When the cook tastes the soup, that's formative assessment.
>
> When the customer tastes the soup, that's summative assessment'

The instructor of an online course can be less formal in formative assessment. Often, in my instruction, formative assessment takes place in individual e-mail discussions with students. These result in written documentation of the student's understanding of a concept. Discussion Boards are also useful for formative assessment. There is a transparency to a student's discussion of his or her understanding of a concept that often doesn't occur in the student's "formal" writing on an essay test. The instructor's participation in Discussion Boards--clarifying misunderstandings in a gentle way--is an excellent example of formative assessment.

Five: Criterion Referenced or Norm Referenced

The direct numerical report of a student's test performance is the student's raw score, that is, the number of correct answers. Most often, we cannot interpret raw test scores as we do physical measures such as height because raw scores have no true meaning. Therefore, the way we can meaningfully talk about test scores is to bring in a referent. There are two major referents for tests: norm-referencing and criterion-referencing. The difference between norm- and criterion-referenced tests is their interpretation; that is; how we derive the meaning from a score. Norm-referenced tests are constructed to provide information about the *relative* status of students. Thus, they facilitate comparisons between one student's score to the score distribution, that is, the mean and standard deviation of some norm group.

Alternatively, when student work is scored or graded using a criterion-referenced scoring, the student's work is compared with a "standard" of expected work or is graded using a rubric designed to be a descriptor of expected work. The "standard" or scoring rubric should match the instructor's delineated objectives. Using criterion-referenced scoring in which the student's work is compared with the criterion presented in the assignment, allows all students in the class to obtain "mastery" of the assignment. In contrast with norm-referenced scoring, whereby only a very few students can reach the top score. Criterion-referenced scoring encourages cooperative learning and the sharing of ideas (Klecker 2003, Linn & Miller, 2005). I have posted short multiple-choice quizzes (with knowledge, understanding, and application level questions) along with a "Quiz Discussion Board" and have encouraged students to discuss the questions in order to determine their individual answers. These formative quizzes are simple devices for facilitating learning. When all students can obtain the criterion-referenced "A" sharing knowledge and information comes naturally.

Six: Validity

Content validity is the major measurement issue in assessment in classroom assessment (Linn & Miller 2005). The best way of assuring that the instructor is teaching the material that he or she planned to teach and assessing what he or she taught the construction of a Table of Specifications [also called a Test Blueprint]. Table 1 is an example of a generic Table of Specifications using the cognitive levels presented in Bloom, et al. (1956).

Table 1. *Generic Table of Specifications Using Bloom's Cognitive Levels*

Content	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation
Concept 1	2 (items 8, 13)	-	-	3(items10,13,16)	-	-
Concept 2	3(items3,14 ,17)	-	-	-	-	-
Concept 3	3(items 9,10,18)	2 (items 4 ,11)	3(items 5,12,21)	-	-	-
Concept 4	-	3(items 6,7,15)	-	-	-	-
Concept 5	-	-	2(items 9,20)	-	1 (item 22)	1 (item 23)

The six cognitive levels of Bloom's taxonomy are across the column headings in order of ascending complexity (Table 1). Five concepts (featured in the objectives of the course) are listed under "Content." The number of items that the instructor plans to use to measure the emphasized concept at the appropriate cognitive level is delineated in the Table of Specifications. The test (in this case a 23-item multiple-choice and short-essay test) items will mirror the plan. Thus, the instructor is assured that he or she has matched the test items to the course objectives and has assured the content validity of the test.

Seven: Reliability

Many measurement scholars argue that reliability is secondary to content validity in classroom assessment (e.g., Brookhart 2000, Linn & Miller 2005, Sax 1997). This argument

concerns the strength of the reliability coefficient. However, it does not extend to the requirement of objective scoring, the essence of reliability. Objective scoring is defined as scoring that is planned and carried out so that two or more scorers would obtain the same score. This is most easily done with multiple-choice questions. It is less easily done with restricted response or essay items. For the latter, clear rubrics should be designed and applied conscientiously. Providing the scoring rubrics to students at the time the assignment is given takes the mystery out of the assignment and places the control over learning into the student's hands. Making learning targets clear is always a good idea. Another facet of reliability is the students' ability to guess on multiple-choice items. As the instructor is trying to measure what the student truly knows, guessing should be minimized. For this reason, true-false items should not be used in classroom assessment as the student has a 50/50 chance of getting a correct answer by mere guessing.

Eight: Diversity

Following ADA guidelines for online classes to ensure equal access is the *sine qua non* of this section (see Edmonds 2003). These guidelines should be integrated into the design of the course and *must* be considered when writing the learning objectives for the course. Beyond this, the often cited--if not measurable--"learning style" differences should be given consideration when designing online assessment. Some students will excel at responding to multiple-choice exams (with immediate feedback), others will prefer short essay items (scored with rubrics), still others prefer lengthy written assignments such as term papers. Other students will prefer more creative assessments such as cognitive mapping and web-quests. A number of varied assessments will allow for the multiple measures required for reliability and will make

assessment an integrated, fun part of the course. Assessment should be as "multivaried" as your students!

Nine: Variety

The variety of assessment strategies available to the online professor is very appealing. There is a wide variety of material that lends itself to case studies. Case studies can be considered in a Discussion Board format for formative assessment--with discussion from other students with guidance from the instructor. Similar case studies can then be used for summative assessment with the addition of multiple-choice and short-essay questions. Problem-solving scenarios can be used for both teaching and assessment by using the Discussion Board. Forming small groups within an online class enables group projects and presentations. Allowing students to discuss test questions before responding as individuals leads to greater learning (Klecker, 2003). Having students develop web-quests, construct concept maps, and solve unique problems not only adds to the variety of assessments, but decreases opportunities to plagiarize that are present in the assignment of "term papers" (Bauer & Anderson 2000).

Ten: Providing Feedback

The importance of prompt (if not immediate) feedback is important in online classroom assessment (Cashion & Palmieri 2002, Siew 2003, Shuey 2002). Students in face-to-face classrooms expect graded work to be returned within the week. Because the time parameters are different in online classes, feedback to students can range from instant--for example in a Blackboard-graded multiple-choice exam--to weekly--as in an instructor-graded essay exam. Feedback to students serves as both an extrinsic motivator--when grades are involved--and an intrinsic motivator--when self-correcting is the primary motivating force.

Feedback comes in two major categories: formative and summative as discussed in number four above. I often encourage students to submit assignments in progress for feedback. This works well online for term papers and research proposals. (This tactic also makes the summative evaluation a lot easier!) Greenberg (1998) delineated the features of online assessment that lead to instant gratification for students in the form of instant feedback. Further, the immediate scoring and "Grade Book" features provide instant gratification for instructors as well.

References

Bauer, J. F., & Anderson, R. S. (2000). Evaluating students' written performance in the online classroom. *New Directions for Teaching and Learning, 84*, 65-72.

Benson, A. D. (2003). Assessing participant learning in online environments. New Directions for Adult and Continuing Education, 100, 69-77.

Bloom, B.S., Englehart, M.D., Frost, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives. Handbook I: Cognitive domain*. New York, NY: David McCoy.

Brookhart, S.M. (2004). *Grading*. Upper Saddle River, NJ: Pearson, Merrill, Prentice-Hall.

Brookhart, S. M. (2000). *The art and science of classroom assessment: The missing part of pedagogy*. ERIC Digest. ERIC Clearinghouse on Higher Education, One Dupont Circle, Washington, DC (ERIC Document Reproduction Service No. ED432938)

Cashion, J., & Palmieri, P. (2002). *The secret is the teacher: The learners' view of online learning*. Leabrook, Australia: National Center for Vocational Education Research (ERIC Document Reproduction Service No. ED475001)

Edmonds, C. D. (2003). Information technology access: State, federal and international law. U. S. Department of education, Office of Postsecondary Education. Retrieved January 1, 2005 from http://conference.merlot.org/conference/2003/presentations/MIC03_Edmonds.access.ppt.

Greenberg, R. (1998). Online testing. *Techniques: Making Education & Career Connections, 73*, (3), 109-131.

Gronlund, N.E. (2004). *How to write and use instructional objectives* (5th ed.). Upper Saddle River, NJ: Pearson Merrill Prentice Hall.

Klecker, B. (2003). Formative classroom assessment using cooperative groups: Vygotsky and random assignment, *Journal of Instructional Psychology*, *30* (3), 216-219.

Linn, R. L., & Miller, M.D. (2005*). Measurement and assessment in teaching* (9th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Northcote, M. (2002). Online assessment: foe or fix? *British Journal of Educational Technology 33*, (5), 623-625.

Ricketts, C., &Wilks, S.J. (2002). Improving student performance through computer-based assessment: Insights from recent research. *Assessment & Evaluation in Higher Education, 27*, (5), 475-479.

Sax, G. (1997). *Principles of educational and psychological measurement and evaluation* (4th Ed.). Belmont, CA: Wadsworth Publishing Company.

Shuey, S. (2002). Assessing online learning in higher education. *Journal of Instruction Delivery Systems, 16*, (2), 13-18.

Siew, P. F. (2003). Flexible on-line assessment and feedback for teaching linear algebra. International *Journal of Mathematical Education in Science and Technology, 34*, (1), 43-51. Retrieved October 6, 2004 from http://www.tandf.co.uk/journals

Schneider, S. (2003). An ethical minefield. *Training, 40*, (9), 62.