A CRASH COURSE IN TESTING AND ASSESSMENT

G. Donald Allen
Department of Mathematics
Texas A&M University
College Station, TX 77883-3368
dallen@math.tamu.edu

All about Assessment

Assessment is a topic that too many of us know too little about. It forms a cornerstone of modern education and living at all levels. Whether in the high-stakes arena of end-of-year tests of knowledge and skills, in the evaluation of programs and projects, or the measurement of preference and opinion, assessment swirls around us and deeply affects almost everything we do. This note outlines some of the key facets of assessment as applied to education, and is meant only to be a brief introduction, not a handbook. See the references for further discussion and examples.

Basics Assessment

Assessment is one of those hot-button topics these days that attracts the attention of not only school boards and college deans, but also of the Congress and the President. With apparently too many underachieving students resulting in demands for schools accountability, there has resulted state-wide high stakes testing that has inspired the scrutiny of academics, educators and politicians alike. While this paper is not about high stakes testing *per se*, its conclusions do apply to our own teaching when we understand that our exams and final exams are often high stakes tests in the eyes of many students. All of us should become more aware of the nature of assessment and be prepared examine whether our testing methods meet current expectations of the six key aspects of testing: reliability, validity, objectivity, balance, fairness, and practicality.

We all have a colleague that believes he or she can produce the ultimate examination of just a single question. This question, it is claimed, can discriminate A from B students, B from C students, and the rest. What is the case is our single-question inspired colleague contravenes just about every principle of testing there is, and most certainly the six key aspects above.

Before reviewing these six attributes, we consider the purposes of a test are and compare them with the more general goals of assessment. Typically the test is to measure whether or not a math student understands the theory and can solve problems similar to those we've been teaching. Tests are merely forms of assessment. *Assessment* is used to provide information about

- Skills in a content area
- Guide instruction for the individual and for groups
- Provide feedback or information
- Plan further instruction

- Improve instructional practices
- Focus also on what the student can do
- Guide decisions about instruction (Formative)

On the other hand, the purpose of a *test* is to

- Provide feedback about skills learned or obtained via instruction
- Determine the areas in which an individual needs re-teaching
- Provide grades
- Focus on what the student cannot do
- Indicate attainment of skills (Summative)

In light of these guidelines, the question as to how we use tests beyond merely giving grades is worth consideration. For example, how often do we use the test results as a self-check on how well we taught a particular unit? How often are tests used in any formative manner?

Tests and Testing

Educational measurements such as the tests we give have fundamental flaws. How many of us have reflected that merely the wording of a problem on one of our test can mislead a significant fraction of our class, and that student confusion on such a problem can influence performance on other problems? Probably we all have. We customarily adjust and/or curve test scores to account for such problems and also errors in the problems themselves from misjudgment of difficulty to typos to vaguely stated problem objectives. In fact, these are only a few many issues confronting the test maker. It is generally accepted that educational measures are indirect, incomplete, relative, and used as classification (e.g. grades). Most importantly, educational measures and instruments have inherent errors. Among the foremost are questions with unclear purpose or focus, and questions with unfair or misleading distracters.

In the broad perspective, the goals for making tests are to accurately measure our student's performance in such a way that the grades you assign are consistent with student's attainment of the stated goals for the course. However, constructing tests is a topic to which many of us have given insufficient attention. Too often, we simply modify last years' exams, this giving us a sense of correctness, fairness, and consistency. Lost in this legacy approach to testing are the established and essentially self-explanatory goals for making tests: be concise, realistic, and uni-dimensional; use definite terms and clearly delineate expected behavior. The *uni-dimensionality* means that only one dimension or factor should be tested in a question. For example, a problem with a complex reading passage to understand the problem tests two dimensions, reading and problem solving. A question that requires the solution of one problem prior to solving another can be of two or even more dimensions. It is not the intent here to argue that multipart problems should be abolished, but only to note the possible conflict with established fundamentals of testing.

The function of a test is twofold. It can be for advancement and therefore must be criterion referenced, or for achievement of objectives and therefore norm-referenced. It can be both. At Texas A&M for example, we give common semester examinations to all engineering calculus I and II students. Our goals are dual. The first is to gauge whether the student understands the syllabus material. The second is to assure a relative comparison between sections. (The variable factor here is individual grading.) All exam averages are reported by the dozen or so instructors and a common curve is assigned to the exam. So, effectively there are strict criteria but also there is a relative norm referencing mechanism. The practice of giving common exams has proved beneficial in several ways. First, it has drastically reduced the number of complaints about unfairness. Second, the test is usual carefully examined by a committee of several instructors for clarity, correctness, and difficulty. Additionally, there are rarely any typos in the final product. Finally, instructors are compelled to pay heed to the syllabus guaranteeing that in the next course almost all students will have seen the prerequisite material.

What we see in test construction varies across the faculty to a wide degree. From the traditional drill, kill, and challenge problems to the more exploratory problems of the reform movement, Bloom's taxonomy allows for a hierarchy of levels of comprehension.

Simplified Bloom's for testing

- Knowledge (basic)
- Understanding – ability to restate, interpret
- Applications – beyond restatement to include analysis, synthesis, and evaluation

It is instructive to make questions on the same theme or topic that test each of these levels. Often assessing depth of understanding is difficult and perhaps one method for such a determination is a test construction consisting of triplets of questions at these levels over select key topics. Similar to the Bloom's classification is the so-called *cognitive testing* model. It has five attributes: recall, recognition, differentiation, constructed response, and problem solving.

The Six Key Criteria for Testing

The above six key criteria for testing, reliability, validity, objectivity, balance, fairness, and practicality, are fundamental to the test construction and administration processes. Reliability is an aspect of test scores. Suppose we are given a large test bank of questions pertaining to a single variable to be learned (e.g. differentiation rules), and suppose two sets of items are drawn randomly from this bank. A test is *consistent* if students scoring high on one of the question sets will also score high on the other. The degree of consistency among the test scores is called reliability. A test can be reliable but not valid. For example, testing Green's theorem in a College Algebra course will be reliable: everyone will fail, but it isn't relevant. A classroom test should be consistent and relevant, and this combination is called *validity*. Test scores should be as free as possible from artifacts generated by factors other than those being tested. Writing skills, wording idiosyncrasies, time of scoring, examination time limits all figure as extraneous artifacts reducing test *objectivity*.

As indicated above, the measure validity is a function testing relevance to the content area being studied. *Balance* refers to the distributions of question within the content area. For example, in the testing the three forms of linear equations, point-slope, two point form, and slope-intercept, most of the questions come from just one or two of the types, the test may be said to not have balance. The all encompassing category of *fairness*, which may appear to embrace the other points, refers mostly to the fundamental contract of the teacher with the student, that being not to trick or cheat. For example to ask whether the equation $y = 2x + 3$ is or is not quadratic is a likely candidate because technically the answer is 'yes' since linear functions are contained in the set of quadratics. However, many students would opt for the 'no' response for obvious reasons, and justifiably so. Trick questions, teacher biases, and scoring quirks are all unfair. *Practicality* refers to construction, administration, and scoring of the exam. All must be within the reasonable scope of both teacher and student. Too short or too long of a test can be equally impractical. Likewise, questions requiring overly complex responses can be as well.

Assessment

Beyond testing, there are other forms of assessment not used widely in the mathematics community. The general list of all forms includes

- Collaborative/group projects
- Direct observation
- Essays
- Multiple-choice tests
- Oral questioning after observation
- Performance projects
- Portfolios
- Presentations
- Problem sheets
- Projects
- Self-assessment
- Case-studies
- Short-answer questions
- Practical projects

Most are entirely self-explanatory. A few of these taken from a general list, such as "Performance projects" are not as well suited to mathematics courses, but for pre-service teacher preparation courses, for example, these could be significant.

With assessment we must define the capacity and provisions. For capacity, we can measure ability, achievement, personality, skill, opinion, and attitude. In the teaching of mathematics we tend to assess ability and skill, to some balance, while the student questionnaire at the end of the term measures substantially opinion.

Considerable knowledge accrues by assessing student learning of student on a daily basis. There are two simple and effective instruments for this. On a single sheet of paper, ask your students in the final minutes of the paper either of the following.

1. The Muddiest Point
    - What point was least clear during this session?
    - What point needs further clarification?
    - What point prevented me from learning fully the contents of this session?
2. The one minute paper

- What was the most useful or meaningful thing you learned today?
- What questions remain uppermost in your mind as we end this session?

Either can tell you if you are connecting with class and whether some material should be re-taught. Importantly, students will need to articulate and answer by actually recalling what you've done in class that day. Don't worry about responses. Students tend not to be shy when asked their opinions.

The outcomes available from assessment are

1. Provide to students a chance to reflect on what they have learned and need to learn and to teachers a chance to reflect on what they have taught well and what needs further attention.
2. Provide feedback to the teacher on the clarity of given assessments and – to teachers about the progress of learning
3. Providing clarity to students on the type of (mathematics) knowledge valued, to students on the type of (mathematics) proficiency valued, and to teachers about the progress of learning.

To a lesser extent, assessment should be a routine part of classroom activity as it will promote regular class attendance and promote multiple solutions and approaches, giving a well rounded picture that allows each student to show his/her strengths.

Assessment Weaknesses

Weaknesses in assessment are as prominent as the strengths. It is almost impossible to avoid them, and certainly impossible unless you are conscious of their entire spectra. First, there are the lack of *content validity* and *scoring economy*. These tie closely to two of the desirable aspects of test construction discussed above. For multiple choice assessments there are other, more subtle weaknesses that can be exposed through item analysis, for example. Additionally, the category of scorer weaknesses in the form of *leniency, severity* and the *halo effect* result in general unfairness of the scoring of assessments. While grading leniency and severity are essentially self-explanatory, the halo effect can be of two forms. In its classical form, the halo effect is simply using previous item grades to unsuitably influence how the next item is graded. Alternatively, for particular students that have done very well or poorly on the previous tests can influence, the scorer's (teacher's) attitude is influenced toward their grading of future tests. In some cases, the scorer's *informed judgment* toward how a problem should be solved may influence their grading of the problem solution obtained by an alternative approach, though equally correct. Not dissimilar to informed judgment is *limited judgment*, which arises from limited ability of the scorer to recognize a correct response. Finally, there are always *extraneous factors*, such as the time of day, handwriting, spelling, grammar, etc that can unduly influence how a problem is graded.

Summary

Assessment is broadly defined and used for a variety of purposes from testing to placement, from teaching evaluation to programmatic evaluation, from determining preferences to determining skills. In short, it is different in different contexts. It is therefore complex, both in scope and intent. Understanding assessment and its proper

uses is a study detailed with numerous factors, and drawing the proper and accurate conclusions requires a careful analysis of the questions, the scoring rubrics, and the scorers themselves.

References – general references on assessment and item analysis.

1. Angelo, Thomas A. and K. Patricia Cross. (1993). *Classroom Assessment Techniques: a handbook for college teachers.* Jossey-Bass: San Francisco.
2. Hambleton, Ronald K., Swaminathan, H., Rogers, H. Jane. (1991). *Item Response Theory*, Sage Publications, Newbery Park.
3. Marshall, Jon Clark, and Hales, Loyde Wesley. (1972). *Essentials of Testing,* Addison-Wesley Publishing Company, Reading.
4. Sledge, Sharon, (2006). All About Assessment, ICTCM Proceedings of the 2004 Annual Meeting.