



## **Two Traditions for Assessing Student Achievement**

Dr. Art Bangert  
Department of Education  
[abangert@montana.edu](mailto:abangert@montana.edu)

*Dr. Bangert is an assistant professor of the Adult and Higher Education graduate program. His teaching-research agenda includes the use of Project Based Instructional Strategies for teaching assessment literacy to preservice teachers, design issues for standards-based assessment systems, and evaluation of on-line courses.*

### **A Historical Perspective on the Use of Standardized Tests**

Until about 1926, most colleges and universities used locally developed essay tests to evaluate the readiness of applicants to undertake and successfully complete collegiate study (Whitney, 1993). In response to the need for a more efficient and standardized method for screening, the College Entrance Examination Board (CEEB) created the Scholastic Aptitude Test (SAT) in efforts to provide college officials with comparable test results for all candidates. As World War II ended and servicemen took advantage of the GI Bill, an even greater need arose to evaluate the readiness of large numbers of applicants for entrance into university systems across the United States. More recently, admissions-test scores have been used extensively to recruit students into specialized programs of study.

The common metric or scale that standardized admissions test scores are derived from provide a way for making direct comparison of scores across large groups of individuals. The capability for making these types of comparisons allows decision makers to select the most highly qualified applicants for their institutions when minimum test scores are the sole standard for admittance. There are many examples of standardized tests whose scores are used to

determine the eligibility of applicants for admission to general higher education programs and other specific professional programs of study. A few of the more familiar admissions tests include: The American College Testing Program (ACT), the Scholastic Aptitude Test (SAT), the Graduate Record Exam (GRE), the Graduate Management Admissions Test (GMAT), the Law School Admissions Test (LSAT), and the Medical College Admissions Test (MCAT). The list goes on.

Standardized tests that are used to make decisions about student admissions are very different from the traditional informal classroom tests that higher education faculty use to assess student knowledge and skills related to important course content. Historically, standardized achievement tests have almost exclusively consisted of multiple choice items. Classroom tests on the other hand include a greater variety of items types including True-False, short answer and performance tasks in addition to traditional multiple choice test questions. Standardized achievement tests differ from classroom tests in the following ways: (1) the nature of the learning outcomes and content measured, (2) the quality of test items, (3) the reliability of tests, (4) procedures for administering and scoring, and (5) the interpretation of scores (Linn & Gronlund, 2000).

### **Learning outcomes and Content**

Standardized achievement tests assess outcomes and curricular content common to most schools across the United States. Test publishers involve content experts (faculty) in the development of subject-specific items for their assessment systems. Tests such as the ACT are designed to evaluate a student's general educational readiness for college coursework by assessing learning outcomes common across many high school courses. However, classroom tests of student achievement measure achievement related to specific course outcomes and are

better suited for formatively assessing student learning. For example, the ACT assesses prerequisite knowledge and skills that a student graduating from high school would be expected to know when enrolling in college-level English, Science, Social Studies, or Mathematics coursework. However, a classroom assessment for a specific college-level Geology course would be limited to measuring learning outcomes that are specific to an in-depth study of the complex principles and skill applications related to earth science rather than broad general science concepts.

### **Quality of Test Items and Reliability**

Standardized assessments produced by commercial publishers typically contain test items that are of high technical quality. The rigorous review process that preliminary forms of standardized achievement tests undergo is primarily responsible for the superior level of item characteristics. Prior to the final production of achievement test batteries, items are pilot-tested with national samples of students. Statistics from these item try-out studies are reviewed to determine the acceptability of items for use in the final test forms. The criteria for item acceptability is based on statistics that are used to evaluate the reliability and validity of individual test questions as well as groups of test items that comprise achievement test battery subtests. For example, the relationship between individual items that comprise a mathematics subtest are examined as well as the relationship of these same mathematics items to all other items comprising subtests for an entire standardized achievement test battery.

Test-retest and internal consistency reliability are most often referred to when evaluating the consistency of standardized achievement test items. Test–retest reliability refers to the stability of test results overtime. That is, the level of agreement between results for two different administrations of the same test for the same individuals over a two or three week time interval.

Internal consistency reliability, on the other hand, refers to the similarity of items within a test or subtest designed to measure one specific skill area such as reading comprehension or mathematics calculation. For example, we would expect that a group of items written to assess algebra skills would exhibit high levels of internal consistency reliability. However, the homogeneity of this algebra subtest would most likely be adversely affected if items were added that required students to answer questions related to social studies that assess skills unrelated to principles and concepts covered in most algebra courses. The test-retest and internal consistency reliability coefficients for most standardized achievement test batteries is commonly found to range from .80 and .95 (Linn and Gronlund, 2000).

The reliability for most classroom tests created to assess specific course content is in most cases unknown. Faculty typically do not conduct reliability studies for individual classroom assessments. However, internal consistency reliability is probably the most important type of consistency measure that faculty should be concerned with when formatively assessing student progress toward specific course learning outcomes. That is, a classroom test created to assess a specific set of knowledge and skills should consist of items that best represent the construct that is being assessed. The test should not include items that students are unfamiliar with or that contains content that was not covered during instruction. Test – retest and internal consistency reliability of classroom assessments can be improved when guidelines for writing test items recommended by measurement experts are taken into consideration. It is important that test items are written in formats that are clear and understandable and are free from grammatical errors and confusing language. Once items are written they should be reviewed for bias to ensure that certain groups of students will not be put at a disadvantage for selecting the correct responses. Lastly, efforts should be made to ensure that items are appropriate for assessing the cognitive

complexity of intended course learning outcomes (Stiggins, 1998). For example, it would be inappropriate to use multiple choice items when evaluating a student's skills for properly using a microscope. A performance assessment task would be a much more appropriate task for assessing this skill. Popham (2002) provides a comprehensive discussion of item writing rules for a variety of item types that can be used to assess student achievement.

### **Administration and Scoring**

Standardized assessment systems include manuals which provide detailed directions to ensure consistent administration and scoring procedures. The specificity of administration and scoring rules provides the basis for score comparability across large student groups of similar grade levels and ages. The test administration guides provide detailed information related to time limits for subtests, acceptable instructions for clarifying questions students have about test items, and procedures for completing answer documents. The only acceptable reason for deviating from standardized test administration procedures would be to make accommodations for individuals with documented disabilities. The provision of scoring keys and guides are important features of standardized tests because they help to reduce errors when student responses are hand-scored. Most large scale assessments such as the ACT or SAT require that student answer documents be sent to the publisher for machine scoring. One of the benefits of machine scoring is the accuracy with which student responses are scored and reported. One disadvantage of scoring student responses electronically is that instructors fail to realize patterns of student errors that are important to recognize when planning instruction.

The purpose of most classroom tests is to assess student progress toward specific learning goals a course is designed to teach. Although instructors attempt to administer their tests to all students in a similar manner, the procedures used are not nearly as consistent as those required

for standardized achievement tests. For example, parts of a classroom tests might be administered a second time to ensure that students who performed poorly during the first administration have mastered the knowledge or skills that they formerly lacked. It is also not unusual for instructors to assess course competencies using “take home tests”. The advantage of this type of assessment format is that students have adequate time to complete complex, project-based tasks. The disadvantage, however, is that there is little or no control on the time individual students spend completing the test let alone who else may be contributing to their performance. These non-standardized procedures would not be permitted with any of the traditional standardized achievement test batteries.

### **Interpretation of Scores**

Standardized achievement tests use a norm-referenced framework for interpreting student performance by comparing it to the performance of a well defined group of other students who have taken the same test (Nitko, 2004). In other words, scores reported by standardized achievement tests indicate where individuals rank in comparison to other individuals, not how many items they were able to answer correctly. For example, let’s say that a high school senior scored at the 90<sup>th</sup> percentile on the overall composite score for the April administration of the ACT. In this case, the 90<sup>th</sup> percentile rank indicates that this senior did as well or just better than 90% of all other seniors from across the United States who participated in the April administration of the ACT. Test results reported in this manner are considered relative interpretations of student performance as compared to the criterion-referenced interpretations used by most classroom tests.

Traditionally, classroom tests describe student performance in terms of some type of descriptive category. Instructors commonly convert raw scores earned on a test to percentages

and assign a grade based on a pre-established criterion (e.g., 95%-100% = A). Performance tasks are forms of classroom assessments that use descriptive categories represented by rubrics to describe student performance. Rubrics are created to explicitly define student progress toward proficiency of learning outcomes that performance tasks are intended to assess. Performance descriptor categories sometimes referred to as performance standards are represented by terms such as basic, nearing-proficiency, proficient or advanced. These descriptive categories are commonly used by standards-based classroom assessments to label student performance. Performance standards unlike norm-referenced scores are aligned to detailed descriptions of varying levels of student performance relative to skills required for mastery of knowledge and skills represented by content standards.

### **Advantages and Disadvantages of Both Testing Procedures**

Both standardized and classroom testing procedures have advantages and disadvantages. Classroom tests are more flexible; can be readily adapted to revisions made in course content and are best suited for assessing student progress toward specific course outcomes. Yet these types of informal assessments would provide little or no information about how students are performing nationally in a specific content area. Standardized tests are better for portraying an individual's general academic achievement or scholastic preparedness to enter college as compared to other students with similar characteristics. However, standardized tests are broad-based and do not provide the type of diagnostic information that faculty require for providing students with specific and corrective feedback related to mastery of specific course competencies.

### **Conclusion**

The appropriateness of testing procedures used in educational settings is based on the intended uses of test results (AERA, APA, NCME, 1999). Faculty interested in formatively

assessing student progress would be best advised to produce classroom assessments that are clearly aligned to learning goals, use items written according to specifications recommended by measurement experts, are appropriate for assessing the cognitive complexity of intended learning outcomes and are evaluated for bias. The high technical quality that characterizes most published standardized achievement tests is accomplished by involving content area specialists in the production of content specific test items, and conducting national pilot studies that provide statistics that psychometricians review to determine item acceptability. Standardized achievement tests are advantageous for making admission decisions that require comparable results for large and diverse groups of individuals. According to Whitney (1993), a combination of both standardized test scores and prior academic record (i.e., high school grades based on classroom assessments) are the best predictors of college success. In summary, information from both testing formats provides good information about a student achievement. When considered collectively both types of assessments offer a much more complete picture of a student's capabilities than when considered in isolation.

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Psychological Association.
- Nitko, A.J. (2004). *Educational assessment of children (4<sup>th</sup> ed.)*. Upper Saddle River, NJ: Pearson.
- Popham, W.J. (2002). *Classroom assessment: What Faculty need to know (3<sup>rd</sup> ed.)*. Boston: Allyn & Bacon.
- Stiggins, R. J. (1998). *Classroom assessment for student success*. National Education Association: Washington, DC.
- Whitney, D.R. (1993). Educational admissions and placement. In R.L. Linn (Ed.). *Educational measurement*. (3<sup>rd</sup> ed.), Phoenix, AZ: The Oryx Press.