Advances in Social Sciences Research Journal – Vol.4, No.21
Publication Date: Nov. 25, 2017
DoI:10.14738/assrj.421.3831.
Allanson, P. B., Rawlings, R., & Notar, C. E. (2017). A Parent and Student Guide to the ABC's of Norm and Criterion Reference Testing. *Advances in Social Sciences Research Journal, (421)* 190-198.

A Parent and Student Guide to the ABC's of Norm and Criterion Reference Testing

Dr. Patricia Bolton Allanson
Liberty University

Dr. Robin Rawlings
Walden University

Dr. Charles E. Notar
Jacksonville State University (ret.)

ABSTRACT

It has been stated that one year of a students primary and secondary education is used for testing. The importance of having a full understanding of the types of standardized tests millions of students will experience can not be understated. Since the mid 1800's, standardized testing has been part of American education. Specifically, educators use norm and criterion referenced testing as part of their profession. Parents and students need to understand what these tests mean, specific differences, and how they bear upon their lives. Although teachers use these tests to make judgments, the students and parents must understand the results of norm and criterion referenced assessments. The following is a guide assembled by the authors to help in that understanding those results. The authors have a combined 100 years of experience in education at the high school, junior college and university level (Reynolds, et al., 2006)..

Key words: Norm Reference Testing, Criterion Reference Testing, Primary and Secondary Education

INTRODUCTION

The *No Child Left Behind Act* (NCLB) of 2002 mandated that all 50 states administer annual tests to measure student performance and progress in the learning. After 13 years of NCLB, the focus of education shifted to college and career readiness as indicated in the terms of the *Every Student Succeeds Act* (ESSA). Various states measure progress by adopting their own curriculum standards for what they consider important for students to know, and administer standardized assessments based on these standards. These assessments are considered tools that provide information on how well students are learning, and what steps can be taken to improve outcomes. According to former president of the American Educational Research Association (AERA), W. James Popham (2005), standardized tests are defined as "any test that's administered, scored, and interpreted in a standard, predetermined manner" (para. 6). The following is a historical perspective of how standardized testing evolved into what it is today, and the what the law requires from each state.

HISTORY OF STANDARDIZED TESTING

In ancient times, informal testing was administered through dialogue or essay formats. Socrates (469-399 BCE) was a Greek philosopher who tested his student through conversation. The outcome of his testing procedure was not to obtain a right or wrong answers, but to lead to higher knowledge. Government job applicants in Imperial China in the A.D. 7th century would submit written essays about Confucian philosophy that resembled a more standardized format.

Technology advances in Europe during the 15th century, such as the invention of the printing press and paper manufacturing, continued to power the use of written exams to assess students.

Education in early America was perceived as didactic where the teacher taught both learning to read and moral ethics. Students of all ages would gather in one room, and were taught by a single teacher. In addition to passing on knowledge, teachers were also responsible for assessing students, in a manner much like Socrates, through oral examinations. As education for the masses became widespread, written assessments became more prevalent. The period between 1875 to the end of World War I broadened the goals of American education which provided for the development of norm-referenced testing instruments to measure mental ability in addition to how well students were prepared for college. Examples of testing instruments during this period included common college entrance exams, proposed by Harvard President Charles William Eliot in 1890; the Stanford-Binet Intelligence Test developed by French psychologist Alfred Binet in 1905; and Army Mental Tests to assess job aptitude of US servicemen during the war. This was also the period in which the College Entrance Examination Board (now called the College Board) was founded (1900), and who played a large role in turning the focus of standardizing testing to measure learning rather than innate intelligence. The College Board developed the Scholastic Aptitude Test (SAT) and was first administered in 1926.

The earliest known standardized multiple-choice test, The Kansas Silent Reading Test (1914-1915) created by Frederick J. Kelly, was developed to reduce the amount of time and effort in test administration. In 1938, a holdover technology instrument which detected electrical current flowing through graphite from No. 2 pencils, was introduced. Marketed until 1963, Reynold B. Johnson, an employee for International Business Machines Corporation (IBM) created the IBM 805 test scoring machine utilizing optical mark recognition (OMR) technology.

In 1965, President Lyndon Johnson enacted the Elementary and Secondary Education Act (ESEA). Johnson's purpose for this law was to improve education in poverty stricken school districts, and provided federal funds as a way to fight the war on poverty. This educational act also included testing and accountability provisions to provide test-based evidence that ESEA dollars were being spent appropriately, however, the data collected from these assessments were not enough to identify how specific groups of students were performing. Since 1965, the ESEA has been reauthorized many times.

The No Child Left Behind Act (NCLB) signed into law by President George W. Bush in 2002 (The ABC's of ESEA and No Child Left Behind, n.d.). The NCLB mandated annual testing in three subject areas, reading, math and science to grades 3 through 8 and then a final assessment in 10th grade. Schools are required to show Adequate Yearly Progress (AYP) or face strict penalties. According to Public Law PL 107-110 (NCLB Act of 2002), individual states were required to provide assessments that were: aligned with the state's academic content standards; given to all students; provide reasonable adaptions and accommodations for students with disabilities; involve multiple measures of students' academic achievement that assess higher-order thinking; and are peer reviewed by the U.S. Department of Education.

Although NCLB was a step in the right direction, in respect to indicating where students were making progress according to content standards, the prescriptive provisions became increasingly difficult to implement. On December 10, 2015, President Obama signed into law the *Every Student Succeeds Act* (ESSA). There are many similarities between NCLB and ESSA, however, the ESSA provides for more state and local flexibility in accordance with

accountability, standards and assessments, annual report cards, federal education programs, teachers and school leaders, and school improvement (National PTA, 2016), (see figure 1).

Accountability	• Eliminated adequate yearly progress (AYP); • States establish long-term goals; • States are required to measure multiple measures of student success.
Standards and Assessment	• States adopt standards to prepare students for college or career; standards must include three levels of performance (cognitive complexity); • Assessments requirements remain the same as NCLB.
Annual Report Cards	• Annual report cards requirements remain the same as NCLB, however includes more requirements on student achievement and school information.
Federal Education Programs	• Developed Student Support and Academic Enrichment grant and Statewide Family Engagement Centers program (reauthorized and renamed from The Parent Information Resource Centers).
Teachers and School Leaders	• Eliminated "highly qualified teacher" requirement and evaluation systems.
School Improvement	• Report every three years support and improvement efforts for lowest-performing Title 1 schools, and underperforming subgroups; • Require districts to develop School Improvement Plans in partnership with parents.

Figure 1. Every Student Succeeds Act Highlights

TYPES OF ASSESSMENTS

Assessments are only one piece of the learning process. The learning process itself can be viewed as a triangular cycle including the three aspects of Pedagogy, Content, and Assessment (Harlen, 2014), (see figure 1). All three areas are considered the "curriculum" which a student will experiences throughout their school career. In the classroom setting, students are exposed to an array of assessment formats from informal (also known as formative assessments, or low-stakes tests such as teacher made, observations, etc.) to formal assessments.

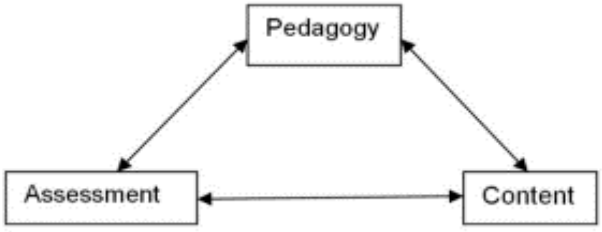

Figure 1. Interactions among aspects of the whole curriculum

There are two types of assessments: formative and summative. Formative assessments are an ongoing process where information is gathered on a students' progress with short term goals

(Harlen, 2014). These types of assessment are used to drive instruction as teachers will adjust lessons based on results. Feedback is provided to students in order to identify the areas in which improvement is needed or what steps the students need to take next. In using formative assessments, the students can also have a role in their own learning by reflecting on their efforts and results of the learning activities. These types of assessments do not have grades or other means of differentation between indviduals attached to them.

Summative assessments, also known as formal, end of unit, semester tests, or high-stakes standardize testing, are the typical type of assessments schools and states administer. The federal government and /or states mandate them. There are several key features of this type of assessment:

- administered to gain an overall perspective of student learning for specific learning goals, over a specific period of time, therefore not an everyday occurrence;
- used to collect data for school evaluations and improvement providing for accountability;
- used to monitor progress as compared to other groups of students, or sub-groups making the same progress.

Since high-stakes tests are required by statute, they are viewed as involuntary, however, they do support learning with long term goals. There are two types of high-stakes standardized testing used throughout the United States, Norm Reference Testing (NRT) and Criterion Referenced Testing (CRT), and they will be the focus of the remainder of this discussion.

Definitions of Norm Referenced Testing (NRT) Criterion Referenced Testing (CRT)

Norm Referenced Testing (NRT) A test designed to provide a measure of performance that is interpretable in terms of an individual's relative standing in some known group. **Norm-referenced** refers to standardized tests that are designed to compare and rank test takers in relation to one another. Norm-referenced tests report whether test takers performed better or worse than a hypothetical average student, which is determined by comparing scores against the performance results of a statistically selected group of test takers, typically of the same age or grade level, who have already taken the exam (http://edglossary.org/norm-referenced-test/).

Norm-reference tests are used to determine differences in testing of those in similarity that have previously taken the test, usually by grade and age. These standardized tests are usually explained by percentile or percentages in ranking. As an example, if someone scores in the eightieth percentile did as well or better than eighty percent of other people taking the same test that were of the same grade and age, then this calculates that twenty percent did better. Also, IQ tests are placed in the category of Norm-reference testing, and is often used for placement in special-education, identifying disabilities in autism, dyslexia, and verbal and nonverbal disability. Norm reference testing also encompasses early childhood kindergarten developmental preparedness, in oral-language, cognitive ability, and social learning. These test questions are usually multiple-choice, short answer, and open-ended. Mentioning a few tests include: Stanford Achievement Test, Iowa Test of Basic Skills, TerraNova, and the California Achievement Test (Lok, B., McNaught, C., & Young, K. 2016).

It should be noted that norm-referenced tests cannot measure the learning achievement or progress of an entire group of students, but only the relative performance of individuals within a norm reference (http://edglossary.org/norm-referenced-test/). Simply put the test shows an individuals standing within the group that was tested. Individual student performances are rated and compared the score to other students, thus, using the bell curve (see figure 2). The "bell curve" is shaped like a bell showing the number of students that do poorly, then the larger

group in the middle that does well, then a few at the bottom doing very well. This test is designed to show the specific set of questions are answered differently by comparing students, not checking to see if they learned the material (Lok, et al., 2016). It simply states that the answer of student A may be worse that student B and better than student C. The assessment will differentiate the best students and some cases use criteria for future testing to weed out the best students from the poor students. Some tests like the ACT/SAT will use both norm reference testing and criterion to determine group ranking and individual scores. When creating bell curve questions, they are selected to determine dissimilarity among those taking the test, and not to determine learned information.

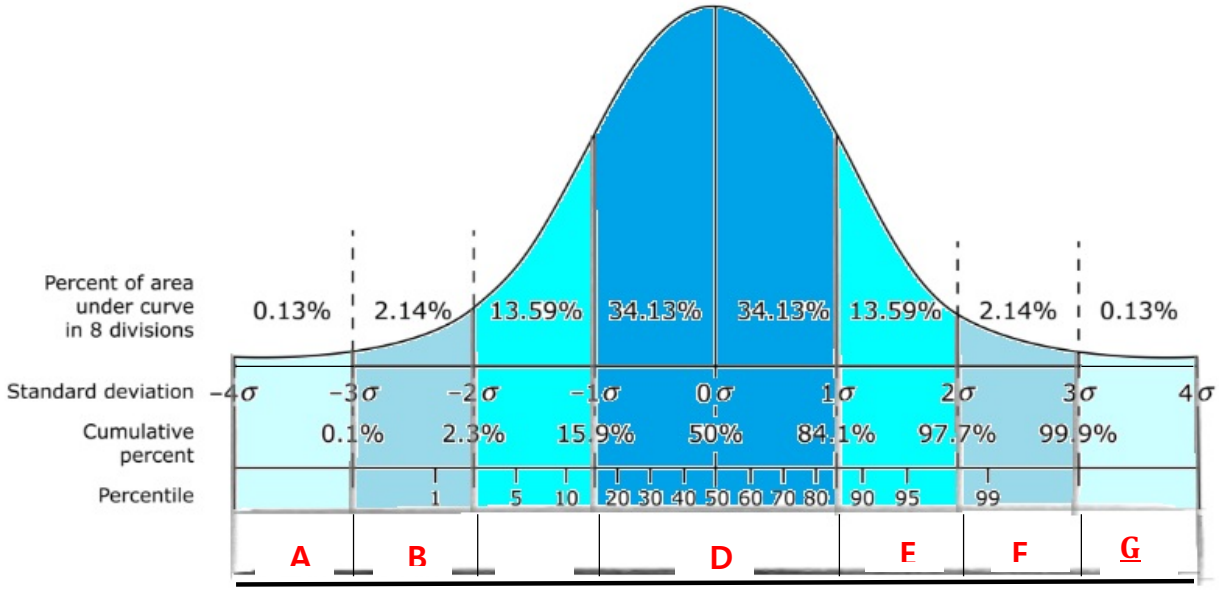

Figure 2: Interpretation of the Bell curve in relation to student skills in each level as compared to other students of the same age/grade level:

A: Skills are extremely below level
B: Skills are significantly below level
C: Skills are below level
D: Skills are average and similar (68%)
E: Skills are above average
F: Skills are significantly above average
G: Skills are extremely above average

Note: The levels describe how a student scores in comparison to other students, however, do not designate specific skill sets.

Figure 2: Bell shaped curve

Criterion Referenced Testing (CRT) Criterion-Referenced test is designed to provide a measure of performance that is interpretable in terms of a clearly defined and delimited domain of learning tasks (https://quizlet.com/78744296/wgu-educational-assessment-flash-cards/). The term criterion-referenced test simply means that the items on the test are referenced to or drawn from a carefully specified set of subordinate skills that make up the goal. On a domain-referenced test, the test items are referenced to or drawn from a carefully delineated domain of tasks. Thus, students' performance on such tests is referenced to the criterion set of skills or domain. Criterion-reference-measures group student performances and are rated to develop a statement about the performance ability of the student.

Oregon State University provides an excellent succinct description of criterion referenced instruction and testing.

Criterion-referenced approach to assessment is designed specifically to synchronize with the particular goals of the course being taught. In some assessment situations, the instructor may want to use a set of clearly stated criteria for evaluation. Criteria usually outlines how a student can reach "mastery" of the identified outcomes and ultimately reach the goals of the contextual lesson.
Recently criterion referenced assessment has taken on the form of scoring guides or rubrics. A scoring guide clearly establishes and describes the specific levels of achievement. The significant difference between criterion referenced assessment and "grades" is that grades are norm-referenced and scoring guides are criterion-referenced. Where we assume that students understand what an A, B, C, D, or F means based on their interpretation, grades by themselves do not describe what a "good" project, performance, and process are. Where criterion-referenced scoring systems overtly describes what "good" ones are. This criteria allows students to work toward mastery of learning tasks
(http://oregonstate.edu/instruction/ed555/zone5/crit.htm).

Cautions When Making Criterion-Referenced Interpretations of Standardized Tests

1. Are the achievement domains (objectives or content clusters) homogeneous, delimited, and clearly specified? If not, avoid specific descriptive statements.
2. Are there enough items (say 10) for each type of interpretation? If not, make tentative judgments and/or combine items into larger content clusters for interpretation.
3. In constructing the test, were the easy items omitted to in- crease discrimination among individuals? If so, remember that the descriptions of what low achievers "can do" will be severely limited.
4. Does the test use selection-type items only? If so, keep in mind that a proportion of correct answers may be based on guessing (this is especially crucial when only a few items are used to measure a specific content domain).
5. Do the test items provide a directly relevant measure of the objectives? If not, base interpretation on what the items actually measured (e.g., "ability to identify misspelled words" rather than "ability to spell." They are related but are not the same process).

Other terms that are less often used but have meanings similar to criterion referenced: objective-referenced, content referenced, domain referenced, and universe referenced.

Differences between NRT with CRT testing Criterion-referenced tests (CRT) and norm-referenced tests (NRT) differ in terms of their purpose and technical characteristics. CRT is considered useful for assessing mastery learning and decision making about instructional change. On the other hand, NRT focuses on producing a ranked ordering of students for specific areas of achievement within a population regarding specific areas of achievement from high achievers to low achievers. Figure 3 shows differences between NRT with CRT testing.

Norm Referenced Testing (NRT)	Criterion Referenced Testing (CRT)
Learning outcomes are described in general or specific terms	Learning outcomes are described in specific terms
Used for survey testing	Used for mastery testing
Measures the individual	Measures proficiency of tasks that students can perform
50% -average of students with correct item response	80%- average of students with correct item response
Compares student performance to other students	Compares student's performance to mastery of curriculum standards
Content covers many objectives (broad)	Content covers few objectives (narrow)
Content sampled Comprehensiveness – one to two items per objective	Content sampled Comprehensiveness – three or more items per objective
Test Plan uses table of specifications	Test Plan focuses on a set of learning tasks
Provides for more variability of scores	Minimal variability
High reliability in relation to the nature of high variability	Estimation of reliability is not appropriate due to limited score variability
Test items are constructed to promote variance or spread. Avoids items that are too easy or too hard.	Test items are constructed to describe performance and relevant responses
Variety of test items used (i.e. multiple choice)	Low dependence on selection of test items.
Emphasis on items that discriminate among students	Emphasis on items that describe student performance
Levels of performance based on rankings	Levels of performance based on absolute standards showing mastery
Interpreting results is based on students' relative standing compared to other students	Interpreting results is based on student performance in relation to achievement domains
Reports results using percentile ranks and standard scores	Reports results using proficiency ranges for failing or acceptable performance

Figure 3: Differences between NRT and CRT Testing

Criterion-referenced tests (CRT) and norm-referenced tests (NRT) differ in terms of their purpose and technical characteristics. CRT is considered useful for assessing mastery learning and decision making about instructional change. On the other hand, NRT focuses on producing a ranked ordering of students for specific areas of achievement within a population regarding specific areas of achievement from high achievers to low achievers. Criterion-referenced and norm-referenced assessment should be seen by parents and students in the what outcomes the assessment has on the students progress in learning.

Important questions parents should ask about testing

One way to learn about the different types of assessments your child will experience throughout the school year, and the implications of the results is to meet with his or her teachers, counselors, or administrators. Bringing a list of questions, such as the one below, will help in this endeavor.

- What types of tests are administered in my child's grade level?
- What are the standards/objectives that my child will be tested on?

- How will my child be prepared for each test?
- What can I do at home to help my child prepare for testing?
- How are the results of each test used (promotion, placement, achievement)?
- What are some strategies used to alleviate test anxiety?

CONCLUSION

Testing, in its many forms, is an accepted occurrence in a students' life. The basics provided in this article on the different types of testing and/or assessments, specifically norm and criterion referenced testing, should be common knowledge that teachers, students, parents, and the general population must understand the measurements used every day to make decisions that impact their lives. It is the authors' hope that parents and students will be able to take this information and use it to ultimately improve their personal learning experiences and achievement.

BIBLIOGRAPHY

Aviles, C. B. (2001). *Grading with norm-referenced or criterion-referenced measurement: To curve or not to curve, that is the question*. ED448403

Background of the Issues [Webpage]. (2016). Retrieved from http://standardizedtests.procon.org/view.resource.php?resourceID=006521

Behuniak, P., & Tucker, C. (1992). The potential of criterion-referenced tests with projected norms. *Applied Measurement in Education, 5*(4), 337-353.

Blackwell Publishers. (1998). Reliability, validity and criterion-referencing. *Journal of Philosophy of Education, 32*(1), 123.

Bloom, B.S. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals.* New York: London.

Burton, K. (2015). Continuing my journey on designing and refining criterion-referenced assessment rubrics. *Journal of Learning Design, 8*(3), 1-13.

Carey, L. M. (1994). *Measuring and evaluating school learning.* (2nd ed.). Boston: Allyn & Bacon, 77.

Dickinson, D. J. (1990). The relationship between ratings of teacher performance and student learning. *Contemporary Educational Psychology*, 15, 142-151.

Glaser, R. (1963). Instructional technology and the measurement of learning out-comes: Some questions. *American Psychologist*, 18, 519–521.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48*(1), 1–47.

Hambleton, R. K., Zenisky, A. L., & Popham, W. J. (2016). **Criterion**-referenced testing: Advances over 40 years. *Educational measurement: From foundations to future.* Wells, Craig S., (Ed); Faulkner-Bond, Molly, (Ed); pp. 23-37; New York, NY, US: Guilford Press; 2016. xvii, 494 pp.

Harlen, W. (2014). *Assessment, standards and quality of learning in primary education*. York: Cambridge Primary Review Trust.

Hart, K. E., & Sciutto, M. J. (1996). Criterion-referenced measurement of instructional impact on cognitive outcomes. *Journal of Instructional Psychology, 23*(1), 26-34.

Hively, W. (1974). Introduction to domain-referenced testing. *Educational Technology*, 14, 5–10.

Lok, B., McNaught, C., & Young, K. (2016). Criterion-referenced and norm-referenced assessments: compatibility and complementarity. *Assessment & Evaluation in Higher Education, 41*(3), 450-465.

Ornstein, A. C. (1993). Norm-referenced and criterion-referenced tests: An overview. *NASSP Bulletin, 77*(555), 28–39.

Pei-Hua, Chen, Tzu-Chun, Chen, & Se-Kang, Kim. (2015). *Comparison of three different kinking procedures between norm-referenced test and criterion-referenced test. International Journal of Intelligent Technologies & Applied Statistics, 8*(1), 71-76.

Popham, W. J. (1990). *Modern educational measurement* (2nd Ed.) Englewood Cliffs, N.J.: Prentice Hall.

Popham, W. J. (2005). *Standardized testing fails the exam* [Blogpost]. Retrieved from https://www.edutopia.org/standardized-testing-evaluation-reform

Popham, W. J. (2014). Criterion-referenced measurement: Half a century wasted? *Educational Leadership, 71*(6), 62-66.

Popham, W. J., & Husek, T. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*(1), 1–9.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2006). *Measurement and assessment in education.* Boston: Person, Allyn & Bacon, 80.

The ABC's of ESEA and No Child Left Behind (n.d.). Retrieved from http://educationpost.org/issues/taking-responsibility/esea-reauthorization/abcs-esea-child-left-behind/