FairTest
The National Center for Fair and Open Testing

Norm-Referenced Achievement Tests

Submitted by fairtest on August 17, 2007 - 1:20pm

Human beings make tests. They decide what topics to include on the test, what kinds of questions to ask, and what the correct answers are, as well as how to use test scores. Tests can be made to compare students to each other (norm-referenced tests) or to see whether students have mastered a body of knowledge (criterion or standards-referenced tests). This fact sheet explains what NRTs are, their limitations and flaws, and how they affect schools.

What are norm-referenced tests?

Norm-referenced tests (NRTs) compare a person's score against the scores of a group of people who have already taken the same exam, called the "norming group." When you see scores in the paper which report a school's scores as a percentage -- "the Lincoln school ranked at the 49th percentile" -- or when you see your child's score reported that way -- "Jamal scored at the 63rd percentile" -- the test is usually an NRT.

Most achievement NRTs are multiple-choice tests. Some also include open-ended, short-answer questions. The questions on these tests mainly reflect the content of nationally-used textbooks, not the local curriculum. This means that students may be tested on things your local schools or state education department decided were not so important and therefore were not taught.

Commercial, national, norm-referenced "achievement" tests include the California Achievement Test (CAT); Comprehensive Test of Basic Skills (CTBS), which includes the "Terra Nova"; Iowa Test of Basic Skills (ITBS) and Tests of Academic Proficiency (TAP); Metropolitan Achievement Test (MAT); and Stanford Achievement Test (SAT, not to be confused with the college admissions SAT). "IQ," "cognitive ability," "school readiness," and developmental screening tests are also NRTs.

Creating the bell curve.

NRTs are designed to "rank-order" test takers -- that is, to compare students' scores. A commercial norm-referenced test does not compare all the students who take the test in a given year. Instead, test-makers select a sample from the target student population (say, ninth graders). The test is "normed" on this sample, which is supposed to fairly represent the entire target population (all ninth graders in the nation). Students' scores are then reported in relation to the scores of this "norming" group.

To make comparing easier, testmakers create exams in which the results end up looking at least somewhat like a bell-shaped curve (the "normal" curve, shown in the diagram). Testmakers make the test so that most students will score near the middle, and only a few will score low (the left side of the curve) or high (the right side of the curve).

Scores are usually reported as percentile ranks. The scores range from 1st percentile to 99th percentile, with the average student score set at the 50th percentile. If Jamal scored at the 63rd percentile, it means he scored higher than 63% of the test takers in the norming group. Scores also can be reported as "grade equivalents," "stanines," and "normal curve equivalents."

One more question right or wrong can cause a big change in the student's score. In some cases, having one more correct answer can cause a student's reported percentile score to jump more than ten points. It is very important to know how much difference in the percentile rank would be caused by getting one or two more questions right.

In making an NRT, it is often more important to choose questions that sort people along the curve than it is to make sure that the content covered by the test is adequate. The tests sometimes emphasize small and meaningless differences among testtakers. Since the tests are made to sort students, most of the things everyone knows are not tested. Questions may be obscure or tricky, in order to help rank order the testtakers.

Tests can be biased. Some questions may favor one kind of student or another for reasons that have nothing to do with the subject area being tested. Non-school knowledge that is more commonly learned by middle or upper class children is often included in tests. To help make the bell curve, testmakers usually eliminate questions that students with low overall scores might get right but those with high overall scores get wrong. Thus, most questions which favor minority groups are eliminated.

NRTs usually have to be completed in a time limit. Some students do not finish, even if they know the material. This can be particularly unfair to students whose first language is not English or who have learning disabilities. This "speededness" is one way testmakers sort people out.

How accurate is that test score?

The items on the test are only a sample of the whole subject area. There are often thousands of questions that could be asked, but tests may have just a few dozen questions. A test score is therefore an estimate of how well the student would do if she could be asked all the possible questions.

***All* tests have "measurement error."** No test is perfectly reliable. A score that appears as an absolute number -- say, Jamal's 63 -- really is an estimate. For example, Jamal's "true score" is probably between 56 and 70, but it could be even further off. Sometimes results are reported in "score bands," which show the range within which a test-takers' "true score" probably lies.

There are many other possible causes of measurement error. A student can be having a bad day. Test-taking conditions often are not the same from place to place (they are not adequately "standardized"). Different versions of the same test are in fact not quite exactly the same.

Sub-scores on tests are even less precise. This is mostly because there are often very few items on the sub-test. A score band for a Juanita's math sub-test might show that her score is between the 33rd and 99th percentile because only a handful of questions were asked.

Scores for young children are much less reliable than for older students. This is because young children's moods and attention are more variable. Also, young children develop quickly and unevenly, so even an accurate score today could be wrong next month.

What do score increases mean?

If your child's or your school's score goes up on a norm-referenced test, does that mean she knows more or the school is better? Maybe yes, maybe not. Schools cannot teach everything. They teach some facts, some procedures, some concepts, some skills -- but not others. Often, schools focus most on what is tested and stop teaching many things

that are not tested. When scores go up, it does not mean the students know more, it means they know more of what is on that test.

For example, history achievement test "A" could have a question on Bacon's Rebellion (a rebellion by Black slaves and White indentured servants against the plantation owners in colonial Virginia). Once teachers know Bacon's Rebellion is covered on the exam, they are more likely to teach about it. But if those same students are given history test "B," which does not ask about Bacon's Rebellion but does ask about Shay's Rebellion, which the teacher has not taught, the students will not score as well.

Teaching to the test explains why scores usually go down when a new test is used. A district or state usually uses an NRT for five to ten years. Each year, the score goes up as teachers become familiar with what is on the test. When a new test is used, the scores suddenly drop. The students don't know less, it is just that different things are now being tested.

Can all the children score above average?

Politicians often call for all students to score above the national average. This is not possible.
NRTs are constructed so that half the population is below the mid-point or average score. Expecting all students to be above the fiftieth percentile is like expecting all teams in a basketball league to win more than half their games. However, because the tests are used for years and because schools teach to them, there are times when far more than half the students score above average.

Why use norm-referenced tests?

To compare students, it is often easiest to use a norm-referenced test because they were created to rank test-takers. If there are limited places (such as in a "Gifted and Talented" program) and choices have to be made, it is tempting to use a test constructed to rank students, even if the ranking is not very meaningful and keeps out some qualified children.

NRT's are a quick snapshot of some of the things most people expect students to learn. They are relatively cheap and easy to administer. If they were only used as one additional piece of information and not much importance was put on them, they would not be much of a problem.

The dangers of using norm-referenced tests

Many mistakes can be made by relying on test scores to make educational decisions. Every major maker of NRTs tells schools not to use them as the basis for making decisions about retention, graduation or replacement. *The testmakers know that their tests are not good enough to use that way.*

The testing profession, in its *Standards for Educational and Psychological Measurement*, states, "In elementary or secondary education, a decision or characterization that will have a major impact on a test taker should not automatically be made on the basis of a single test score."

Any one test can only measure a limited part of a subject area or a limited range of important human abilities. A "reading" test may measure only some particular reading "skills," not a full range of the ability to understand and use texts. Multiple-choice math tests can measure skill in computation or solving routine problems, but they are not good for assessing whether students can reason mathematically and apply their knowledge to new, real-world problems.

Most NRTs focus too heavily on memorization and routine procedures. Mutiple-choice and short-answer questions do not measure most knowledge that students need to do well in college, qualify for good jobs, or be active and informed citizens. Tests like these cannot show whether a student can write a research paper, use history to help understand current events, understand the impact of science on society, or debate important issues. They don't test problem-solving, decision-making, judgement, or social skills.

Tests often cause teachers to overemphasize memorization and de-emphasize thinking and application of knowledge. Since the tests are very limited, teaching to them narrows instruction and weakens curriculum. Making test score gains the definition of "improvement" often guarantees that schooling becomes test coaching. As a result, students are deprived of the quality education they deserve.

Norm-referenced tests also can lower academic expectations. NRTs support the idea that learning or intelligence fits a bell curve. If educators believe it, they are more likely to have low expectations of students who score below average.

Schools should not use NRTs

The damage caused by using NRTs is far greater than any possible benefits the tests provide. The main purpose of NRTs is to rank and sort students, not to determine whether students have learned the material they have been taught. They do not measure anywhere near enough of what students should learn. They have very harmful effects on curriculum and instruction. In the end, they provide a distorted view of learning that then causes damage to teaching and learning.

Attachment	Size
norm refrenced tests.pdf (http://www.fairtest.org/sites/default/files/norm refrenced tests.pdf)	497.6 KB

MEASUREMENT AND EVALUATION:
CRITERION- VERSUS NORM-REFERENCED TESTING

Source: Huitt, W. (1996). Measurement and evaluation: Criterion- versus norm-referenced testing. *Educational Psychology Interactive*. Valdosta, GA: Valdosta State University. Retrieved [date], from http://www.edpsycinteractive.org/topics/measeval/crnmref.html

---

Return to: | Measurement & Evaluation | EdPsyc Interactive: Courses |

Many educators and members of the public fail to grasp the distinctions between criterion-referenced and norm-referenced testing. It is common to hear the two types of testing referred to as if they serve the same purposes, or shared the same characteristics. Much confusion can be eliminated if the basic differences are understood.

The following is adapted from: Popham, J. W. (1975). *Educational evaluation*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Dimension	Criterion-Referenced Tests	Norm-Referenced Tests
Purpose	To determine whether each student has achieved specific skills or concepts. To find out how much students know before instruction begins and after it has finished.	To rank each student with respect to the achievement of others in broad areas of knowledge. To discriminate between high and low achievers.
Content	Measures specific skills which make up a designated curriculum. These skills are identified by teachers and curriculum experts. Each skill is expressed as an instructional objective.	Measures broad skill areas sampled from a variety of textbooks, syllabi, and the judgments of curriculum experts.

Item Characteristics	Each skill is tested by at least four items in order to obtain an adequate sample of student performance and to minimize the effect of guessing. The items which test any given skill are parallel in difficulty.	Each skill is usually tested by less than four items. Items vary in difficulty. Items are selected that discriminate between high and low achievers.
Score Interpretation	Each individual is compared with a preset standard for acceptable achievement. The performance of other examinees is irrelevant. A student's score is usually expressed as a percentage. Student achievement is reported for individual skills.	Each individual is compared with other examinees and assigned a score--usually expressed as a percentile, a grade equivalent score, or a stanine. Student achievement is reported for broad skill areas, although some norm-referenced tests do report student achievement for individual skills.

The differences outlined are discussed in many texts on testing. The teacher or administrator who wishes to acquire a more technical knowledge of criterion-referenced test or its norm-referenced counterpart, may find the text from which this material was adapted particularly helpful.

Additional resources:

- Bond, L. (1996). Norm- and criterion-referenced testing. *Practical Assessment, Research & Evaluation, 5*(2). Retrieved September 2002, from http://ericae.net/pare/getvn.asp?v=5&n=2.
- Linn, R. (2000). Assessments and accountability. *ER Online, 29*(2), 4-14. Retrieved September, 2002, from http://www.aera.net/pubs/er/arts/29-02/linn01.htm.
- Sanders, W., & Horn, S. (1995). Educational assessment reassessed: The usefulness of standardized and alternative measures of student achievement as indicators for the assessment of educational outcomes. *Education Policy Analysis Archives, 3*(6). Retrieved September 2002, from http://olam.ed.asu.edu/epaa/v3n6.html.

FairTest

National Center for Fair & Open Testing

NORM-REFERENCED ACHIEVEMENT TESTS

Human beings make tests. They decide what topics to include on the test, what kinds of questions to ask, and what the correct answers are, as well as how to use test scores. Tests can be made to compare students to each other (norm-referenced tests) or to see whether students have mastered a body of knowledge (criterion or standards-referenced tests). This fact sheet explains what NRTs are, their limitations and flaws, and how they affect schools.

What are norm-referenced tests?

Norm-referenced tests (NRTs) compare a person's score against the scores of a group of people who have already taken the same exam, called the "norming group." When you see scores in the paper which report a school's scores as a percentage -- "the Lincoln school ranked at the 49th percentile" -- or when you see your child's score reported that way -- "Jamal scored at the 63rd percentile" -- the test is usually an NRT.

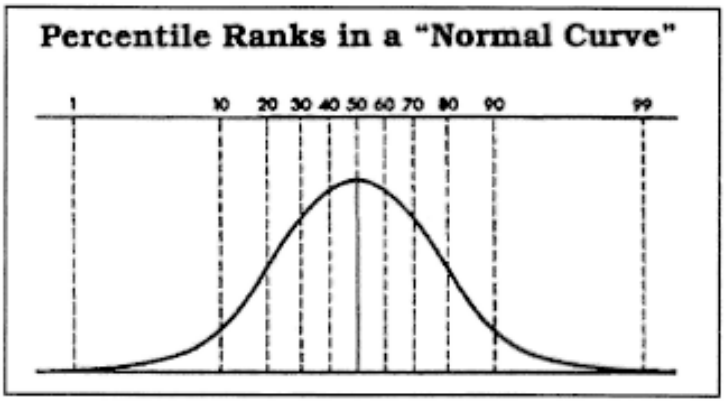
Most achievement NRTs are multiple-choice tests. Some also include open-ended, short-answer questions. The questions on these tests mainly reflect the content of nationally-used textbooks, not the local curriculum. This means that students may be tested on things your local schools or state education department decided were not so important and therefore were not taught.

Commercial, national, norm-referenced "achievement" tests include the California Achievement Test (CAT); Comprehensive Test of Basic Skills (CTBS), which includes the "Terra Nova"; Iowa Test of Basic Skills (ITBS) and Tests of Academic Proficiency (TAP); Metropolitan Achievement Test (MAT); and Stanford Achievement Test (SAT, not to be confused with the college admissions SAT). "IQ," "cognitive ability," "school readiness," and developmental screening tests are also NRTs.

Creating the bell curve

NRTs are designed to "rank-order" test takers -- that is, to compare students' scores. A commercial norm-referenced test does not compare all the students who take the test in a given year. Instead, test-makers select a sample from the target student population (say, ninth graders). The test is "normed" on this sample, which is supposed to fairly represent the entire target population (all ninth graders in the nation). Students' scores are then reported in relation to the scores of this "norming" group.

To make comparing easier, testmakers create exams in which the results end up looking at least somewhat like a bell-shaped curve (the "normal" curve, shown in the diagram). Testmakers make the test so that most sudents will score near the middle, and only a few will score low (the left side of the curve) or high (the right side of the curve).



342 Broadway, Cambridge, Mass. 02139 (617) 864-4810 FAX (617) 497-2224
Web Site: www.Fairtest.org Email: Fairtest@aol.com

Scores are usually reported as percentile ranks. The scores range from 1st percentile to 99th percentile, with the average student score set at the 50th percentile. If Jamal scored at the 63rd percentile, it means he scored higher than 63% of the test takers in the norming group. Scores also can be reported as "grade equivalents," "stanines," and "normal curve equivalents."

One more question right or wrong can cause a big change in the student's score. In some cases, having one more correct answer can cause a student's reported percentile score to jump more than ten points. It is very important to know how much difference in the percentile rank would be caused by getting one or two more questions right.

In making an NRT, it is often more important to choose questions that sort people along the curve than it is to make sure that the content covered by the test is adequate. The tests sometimes emphasize small and meaningless differences among testtakers. Since the tests are made to sort students, most of the things everyone knows are not tested. Questions may be obscure or tricky, in order to help rank order the testtakers.

Tests can be biased. Some questions may favor one kind of student or another for reasons that have nothing to do with the subject area being tested. Non-school knowledge that is more commonly learned by middle or upper class children is often included in tests. To help make the bell curve, testmakers usually eliminate questions that students with low overall scores might get right but those with high overall scores get wrong. Thus, most questions which favor minority groups are eliminated.

NRTs usually have to be completed in a time limit. Some students do not finish, even if they know the material. This can be particularly unfair to students whose first language is not English or who have learning disabilities. This "speededness" is one way testmakers sort people out.

How accurate is that test score?

The items on the test are only a sample of the whole subject area. There are often thousands of questions that could be asked, but tests may have just a few dozen questions. A test score is therefore an estimate of how well the student would do if she could be asked all the possible questions.

***All* tests have "measurement error."** No test is perfectly reliable. A score that appears as an absolute number -- say, Jamal's 63 -- really is an estimate. For example, Jamal's "true score" is probably between 56 and 70, but it could be even further off. Sometimes results are reported in "score bands," which show the range within which a test-takers' "true score" probably lies.

There are many other possible causes of measurement error. A student can be having a bad day. Test-taking conditions often are not the same from place to place (they are not adequately "standardized"). Different versions of the same test are in fact not quite exactly the same.

Sub-scores on tests are even less precise. This is mostly because there are often very few items on the sub-test. A score band for a Juanita's math sub-test might show that her score is between the 33rd and 99th percentile because only a handful of questions were asked.

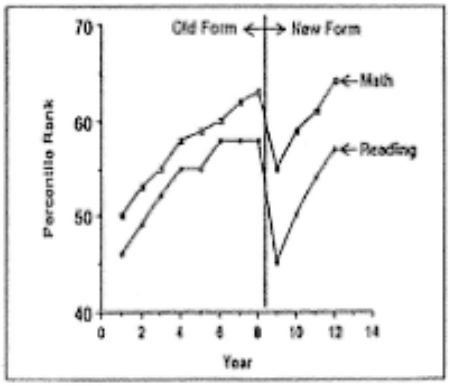
Scores for young children are much less reliable than for older students. This is because young children's moods and attention are more variable. Also, young children develop quickly and unevenly, so even an accurate score today could be wrong next month.

What do score increases mean?

If your child's or your school's score goes up on a norm-referenced test, does that mean she knows more or the school is better? Maybe yes, maybe not. Schools cannot teach everything. They teach some facts, some procedures, some concepts, some skills -- but not others. Often, schools focus most on what is tested and stop teaching many things that are not tested. When scores go up, it does not mean the students know more, it means they know more of what is on that test.

For example, history achievement test "A" could have a question on Bacon's Rebellion (a rebellion by Black slaves and White indentured servants against the plantation owners in colonial Virginia). Once teachers know Bacon's Rebellion is covered on the exam, they are more likely to teach about it. But if those same students are given history test "B," which does not ask about Bacon's Rebellion but does ask about Shay's Rebellion, which the teacher has not taught, the students will not score as well.

Results of Changing to a New Test Form



This chart shows a typical change in test scores when a new test is introduced. Scores go up with one test, down with a new test, then up again. In this example, the reading score was lower with the start of the new test than it was with the start of the old test: all the gain on the old test was false. On the math test, however, some of the gain on the old test appears to be real.

Teaching to the test explains why scores usually go down when a new test is used. A district or state usually uses an NRT for five to ten years. Each year, the score goes up as teachers become familiar with what is on the test. When a new test is used, the scores suddenly drop. The students don't know less, it is just that different things are now being tested.

Can all the children score above average?

Politicians often call for all students to score above the national average. This is not possible. NRTs are constructed so that half the population is below the mid-point or average score. Expecting all students to be above the fiftieth percentile is like expecting all teams in a basketball league to win more than half their games. However, because the tests are used for years and because schools teach to them, there are times when far more than half the students score above average.

Why do schools use norm-referenced tests?

To compare students, it is often easiest to use a norm-referenced test because they were created to rank test-takers. If there are limited places (such as in a "Gifted and Talented" program) and choices have to be made, it is tempting to use a test constructed to rank students, even if the ranking is not very meaningful and keeps out some qualified children.

NRT's are a quick snapshot of some of the things most people expect students to learn. They are relatively cheap and easy to administer. If they were only used as one additional piece of information and not much importance was put on them, they would not be much of a problem.

The dangers of using norm-referenced tests

Many mistakes can be made by relying on test scores to make educational decisions. Every major maker of NRTs tells schools not to use them as the basis for making decisions about retention, graduation or replacement. *The testmakers know that their tests are not good enough to use that way.*

The testing profession, in its *Standards for Educational and Psychological Measurement*, states, "In elementary or secondary education, a decision or characterization that will have a major impact on a test taker should not automatically be made on the basis of a single test score."

Any one test can only measure a limited part of a subject area or a limited range of important human abilities. A "reading" test may measure only some particular reading "skills," not a full range of the ability to understand and use texts. Multiple-choice math tests can measure skill in computation or solving routine problems, but they are not good for assessing whether students can reason mathematically and apply their knowledge to new, real-world problems.

Most NRTs focus too heavily on memorization and routine procedures. Mutiple-choice and short-answer questions do not measure most knowledge that students need to do well in college, qualify for good jobs, or be active and informed citizens. Tests like these cannot show whether a student can write a research paper, use history to help understand current events, understand the impact of science on society, or debate important issues. They don't test problem-solving, decision-making, judgement, or social skills.

Tests often cause teachers to overemphasize memorization and de-emphasize thinking and application of knowledge. Since the tests are very limited, teaching to them narrows instruction and weakens curriculum. Making test score gains the definition of "improvement" often guarantees that schooling becomes test coaching. As a result, students are deprived of the quality education they deserve.

Norm-referenced tests also can lower academic expectations. NRTs support the idea that learning or intelligence fits a bell curve. If educators believe it, they are more likely to have low expectations of students who score below average.

Schools should not use NRTs

The damage caused by using NRTs is far greater than any possible benefits the tests provide. The main purpose of NRTs is to rank and sort students, not to determine whether students have learned the material they have been taught. They do not measure anywhere near enough of what students should learn. They have very harmful effects on curriculum and instruction. In the end, they provide a distorted view of learning that then causes damage to teaching and learning.

For more information contact FairTest, 342 Broadway, Cambridge, MA 02139 (617) 864-4810
FAX (617) 497-2224. Web Site: www.fairtest.org E-mail: FairTest@aol.com