EDUCATIONSECTOR
www.educationsector.org

Explainer

July 2006

Making the Cut: How States Set Passing Scores on Standardized Tests

By Andrew J. Rotherham

ABOUT THE AUTHOR

Andrew J. Rotherham is co-founder and co-director of Education Sector. He also serves on the Commonwealth of Virginia Board of Education. The board's duties include establishing final cut scores on tests for students and teachers in Virginia.

ABOUT EDUCATION SECTOR

Education Sector is an independent education think tank based in Washington, D.C. It is a nonprofit and nonpartisan organization devoted to developing innovative solutions to the nation's most pressing educational problems. The organization seeks to be a dependable source of sound thinking on education policy and an honest broker of evidence in key education debates in Washington and nationally.

ACKNOWLEDGEMENTS

The author thanks Ethan Gray for his research assistance, state officials who shared their expertise, and Chris Cross, Kirk Schroder, and several others who reviewed the report. Special thanks to Education Sector Research Assistant Margaret Price, who provided general research assistance and compiled the state transparency ratings in this Education Sector Explainer.

ABOUT THIS SERIES

Education Sector Explainers give lay readers insights into important aspects of education policymaking. They are not intended to be technical manuals.

1201 Connecticut Ave., N.W., Suite 850, Washington, DC 20036
202.552.2840 • www.educationsector.org

In the current climate of accountability in American public education, tests get more attention and carry more importance than ever before. Both state accountability systems and the federal No Child Left Behind Act hold schools accountable for whether students pass standardized state tests. NCLB requires that schools and school districts make "adequate yearly progress" in reading and math. The law's standard of adequate progress is a sufficient percentage of students passing statewide tests, and it requires serious consequences for schools that continually miss these performance targets.

But states too rarely explain what it actually means for a student to pass a state test, to be "proficient," or how passing scores are established. This gives parents, policymakers and the public only a partial understanding of educational progress and what measures like adequate yearly progress really mean. That's because trying to interpret student performance on a test without understanding the passing score is like reading a map without a scale. This Education Sector Explainer describes the methods states use to set passing or "cut" scores on tests, examines influences on states' score-setting work, and recommends steps to ensure that the public can better understand this important educational process.

Passing scores on state tests are one of three key steps in the development of state academic standards and assessments (Figure 1). First, states must create academic standards. These standards define what students should know and be able to do at different points in their schooling. Then the state develops or purchases tests to assess student progress against these standards. Finally, the state sets passing scores on these assessments, which ultimately determine how demanding the standards and assessments are for students. As a result, understanding the numbers that states release to the public requires an understanding of these underlying decisions.

Figure 1. A Three-Part Process

The cut scores themselves and how they are set are the least discussed of these three steps, even though the cut score codifies what it means for a student to pass or be proficient. In fact, the entire issue of cut scores and the process by which they are set is rarely a focus of much public or media attention at all. This leaves the public with an incomplete picture because understanding the difficulty of passing a test is essential to making sense of student scores, state educational progress, NCLB requirements and various claims and counterclaims about student and school performance.

Setting Cut Scores

On a technical level, states set cut scores along one of two dimensions: The characteristics of the test items or the characteristics of the test takers. It is essential to understand that either way is an inescapably subjective process. Just as academic

standards are ultimately the result of professional judgment rather than absolute truth, there is no "right" way to set cut scores, and different methods have various strengths and weaknesses.[1] The problem is that, though passionate feelings abound, there is no source of agreement about what, for instance, a fifth-grader should know and be able to do in mathematics or what sort of text they should be able to comprehend.[2] The sidebar below describes the nature of the judgments that standard-setters must make.

Though there are many ways to set cut scores, states generally use one of three strategies, Modified Angoff, Bookmarking, or Contrasting Groups.[3]

Judgments

Samuel A. Livingston and Michael J. Zieky describe the judgments standard setters must make:

Any standard—absolute or relative—is based on some type of judgment. A standard is an answer to the question, "How good is good enough?" and this question can only be answered by someone's judgment. The choice of a passing score will involve judgments at some point in the process. It is important that these judgments be:

(1) made by persons who are qualified to make them;
(2) meaningful to the persons who are making them; and
(3) made in a way that takes into account the purposes of the test.

These three requirements are interrelated. Different methods for choosing a passing score require different types of judgments, and, therefore, somewhat different qualifications for the judges.

Source: Samuel A. Livingston and Michael J. Zieky, *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests* (New Jersey: Educational Testing Service, 1982).

Angoff and Modified Angoff

Named after William Angoff who developed the measure in 1971, the Modified Angoff and the Angoff method of setting cut scores depends on judgments about whether a *barely competent* person at a specified level of achievement would be able to answer a particular question correctly.[4] For instance, would a barely "basic" or barely "proficient" student be able to answer a particular math problem on a test? Here's how it works: First, the judges take the actual test students will take and subsequently discuss descriptions of various performance levels for the test, for instance what it means to be proficient, advanced, or basic on it. Then they work through each test item, estimating the proportion of barely basic, barely proficient or barely advanced students that would be able to correctly answer each question. The percentages for each item are then averaged to determine the percentage of all items a test taker would need to answer correctly in order to reach a specific performance level. Table 1 shows this process for a hypothetical test with 10 questions and four judges. (In practice, state tests typically have more than 10 items and more than four judges would be involved in the cut score setting process.)

The judges then see the range of scores their colleagues have selected, discuss the ratings and various performance descriptors for students (for instance what it means to be "proficient"), and then repeat the process a second and even third time, if necessary, to reach a consensus on the cut scores. As a general rule, as the process unfolds the scores tend to converge around the median judgments although this is not formalized in the method.

Table 1. A Hypothetical Ten-Question Test With Four Judges

Question	Judge 1	Judge 2	Judge 3	Judge 4
1	.60	.75	.70	.55
2	.65	.60	.60	.55
3	.65	.65	.60	.50
4	.65	.80	.65	.60
5	.60	.75	.65	.45
6	.55	.80	.60	.45
7	.55	.90	.70	.40
8	.60	.55	.45	.40
9	.55	.65	.50	.55
10	.60	.55	.55	.55
Total	6.0	7.0	6.0	5.0
Average	.60	.70	.60	.50
Cut Score	6	7	6	5

Source: Virginia Department of Education.

The primary difference between the Modified Angoff and Angoff methods is the amount of discretion the judges have in determining the pass-rate probabilities. Under the Angoff method, judges have a theoretical range from zero percent to 100 percent. The Modified Angoff restricts these probabilities to specific numerical choices for the judges, such as 20 percent, 40 percent, 50 percent, 60 percent and 70 percent. Critics argue that specifying probabilities can bias the judges by limiting choices, especially at the high or low end of the spectrum.[5] Other variations of the Angoff Method involve asking judges to estimate whether a barely competent would answer a question correctly and totaling the number of questions each judge answered affirmatively.

Bookmarking

In 1996, CTB-McGraw Hill, an educational publishing and test development company, developed the Bookmaking cut score setting method. The approach has subsequently been used by more than 28 states.[6] Like the Angoff Method, judges review the test and then discuss performance categories. However, under the Bookmarking procedure, the test publisher ranks test questions from easiest to most difficult based on students' past performance on the item, and the judges note or "bookmark" where they think the performance categories should lie along the continuum of test items. This process continues for three rounds, with cut scores ultimately based on the median value of the judges' decisions. During the third round judges are also given data showing the impact of potential cut scores for students taking the test. The sidebar on Page 6 (from the Wisconsin Department of Public Instruction) describes the process in greater detail.

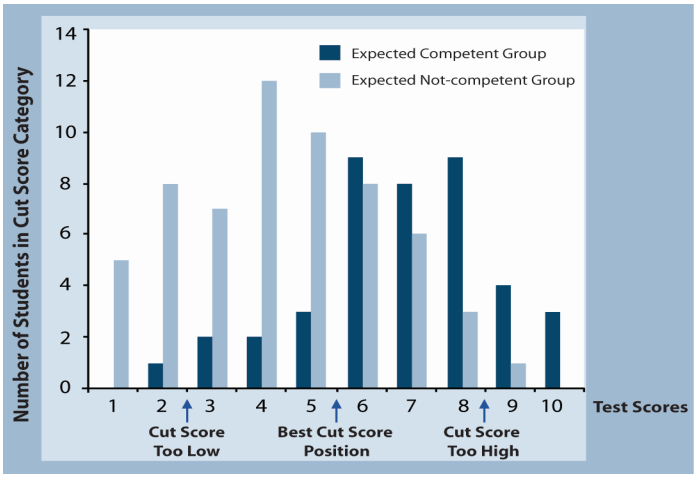
Contrasting Groups

In today's standards-driven environment and the increasing emphasis on tests benchmarked to external criteria, candidate-based cut score methods are less frequent. The Contrasting Groups method is the most common candidate-focused method. This method uses scores from a group of representative test takers to make judgments about where cut scores should be set. Generally, the judges examine a sample of test scores rather than scores from the entire population of test takers. The test takers are delineated by some characteristic known about them, for instance performance on other tests or courses taken, that is deemed to be an accurate way to discriminate amongst them.

Based on this information they are divided into groups based on their expected performance on the test. Their expected performance is compared to their actual performance on the test and based on this information the judges decide on the best cut score to separate the groups. Figure 2 (taken in its entirety from the National Board on Educational Testing and Public Policy's *Results May Vary*) shows the Contrasting Groups method applied to hypothetical test results for two groups of test takers. The scores of the high performers (called the "expected competent group") lay toward the high end of the scale. The scores of the actual test takers (called "expected not-competent group") are toward the low end. In the middle is where the judges typically place the cut score that best delineates the performance category.

Because judges are often working with just a sample of students, the scores of test takers can be wildly uneven. In this case the data are smoothed out statistically to ensure that judges can make their determinations based on a more consistent scale.[7]

Figure 2. Cut-Score Setting Using Contrasting Groups



Source: NBETPP, Monographs Vol. 1, No. 1, April 2000.

What It Looks Like In Practice

The process of score setting can last from a day to several days. Some states run the process with their own staff, while others contract the work to the same companies that developed the test. Participants on score-setting panels usually include teachers and curriculum specialists who teach relevant grade spans. Some states also include teachers from other grades, representatives of higher education and other stakeholders. After the score setting committee reaches its judgment, it is typically ratified by a state's board of education and formally adopted as state policy.

Political considerations can also influence the setting of cut scores—and sometimes do. As a general rule, state policymakers want to look good, and this can create a downward pressure on passing scores. States also often set cut scores lower than they otherwise might in order to create buy-in from educators and the public. While high passing scores might earn plaudits from some educators and school reformers, they can erode public and educator confidence in various reforms because progress appears daunting. Political influences on cut-score setting can be subtle. Decisions about the composition of score-setting panels, for example, can affect the process in largely untraceable but potentially powerful ways.

When interpreting cut scores, it is essential to remember that they are meaningless outside the context of a specific test. So comparing cut scores between tests or states is futile because, for instance, a cut score of 33 out of 50 on one test may or may not be more demanding than a cut score of 27 out of 50 on another test.

If a test has a low cut score, the media and other observers should look closely at the test. But a low cut score may not always be a bad thing. It might be an especially hard test. In the same way, high cut scores don't guarantee rigor. It may be a very easy test. Also, many states convert the student scores on tests into scaled scores, meaning the actual number of questions a student answered correctly is

At the Bookmark Standard-Setting Workshop

Over a three-day period, Bookmark standard-setting participants engage in training, three rounds of structured discussions and ratings where they set cut scores on an assessment, and a writing session where they write descriptions of the recommended performance levels.

Round 1 of Discussion and Bookmark Placement:
In small groups—typically tables of six to eight people—participants examine each item in the OIB [Ordered Item Booklet], discussing what each item measures and what makes the item harder than those before it. After this discussion, each participant determines a cut score by placing a bookmark in the OIB according to his or her own judgment of what, for example, proficient students should know and be able to do.

Rounds 2 and 3 of Discussion and Bookmark Placement:
Participants then engage in two more rounds of bookmark placements. In Round 2, participants discuss the rationale behind their original bookmark placement with other participants at their table. In Round 3, participants at all tables discuss their bookmark placements together. After each round of discussion, participants may adjust or maintain their bookmark placements. Impact data, that is the percentage of students in that state that would fall below each bookmark, is introduced to participants during the third round. After the final round of bookmark placement, CTB Research calculates the group's recommended cut score by taking the median of all bookmark placements in the final round.

Description Writing:
To complete the standard setting, participants write performance-level descriptors that reflect the final recommended cut scores. In small groups, participants examine the items before the bookmark and synthesize the content measured by those items. After two rounds of revisions, the performance-level descriptors represent a summary of the knowledge, skills, and abilities students must be able to demonstrate to enter each performance level.

Second Method Validation:
For some states, a second method of standard-setting can be introduced at the end of the Bookmark standard-setting workshop in order to validate the recommended cut scores. CTB recommends using the contrasted groups method where teachers who have set the cut scores using the Bookmark methods then assign each of the students in their class rosters to a given performance level. CTB can check the actual performance of that student against the teacher's judgment. With cut scores being used for important decisions (e.g., WKCE is one factor in the promotion decision for students in Wisconsin), a second method gives the state more information upon which to set the final cut scores.

Final Decision on Cut Scores:
The final determination of a state's cut scores is made by the individual or group with education policy-making authority within the state. The State Superintendent and/or the State Board of Education considers the recommendations of the standard-setting committee and, oftentimes, review of those recommendations by a Technical Advisory Committee when establishing the final performance levels for a state's assessments. The more information the policymakers have when making this decision, the more defensible the final decision will be if the cut scores are challenged by a school, student or parent.

Source: Wisconsin Department of Public Instruction.

converted to a scale such as that used to report SAT scores. Consequently, when considering cut scores, it's essential to know what the figures represent. Is the state reporting a raw score or a scaled score?

It is also important to remember that cut scores do not equate with traditional letter grades in education. On a difficult test a cut score that represents answering correctly 65 percent of the test items may in fact be much more challenging than "D" work. Conversely, on an easy test a score of 80 percent may not reflect a high level of learning.

To illustrate the outcome of a cut score setting process, Appendix 1 (Page 9) shows the cut scores for South Dakota in grades three through eight and grade 11. The appendix shows the number of reading and math questions students need to answer correctly to pass the state's tests, and it depicts the translation of those results into scale scores.

Appendix 1 reads as follows: For a fourth-grade student to be proficient in reading they would have to answer at least 26 items correct on the test.

Clear Cutting

To find out how much cut score information states make available, Education Sector surveyed all 50 state departments of education. The method was straightforward: We looked for information on the Internet about raw cut scores. States that only posted scaled scores were not counted as being transparent.

We searched each state's Web site.[8] First, we looked for a link for assessment or accountability or a similar heading. Then we searched state assessment manuals or guides. If we could not easily locate this information, we sifted through individual documents related to testing. If we could locate the information on cut scores, the state was rated as "transparent." If we searched thoroughly and did not find the relevant information the state was rated as "non-transparent." Table 2 shows our results.

Table 2. State Transparency Ratings: Is Information About Cut Scores Available on the Internet?

Transparent (Information Available on the Web)			Non-Transparent (Very Difficult To Locate/ Unavailable On The Web)	
Alaska	Maine	Ohio	Alabama	Nebraska
Arizona	Maryland	Oregon	Florida	Nevada
Arkansas	Massachusetts	Pennsylvania	Georgia	North Carolina
California	Minnesota	South Carolina	Hawaii	North Dakota
Colorado	Mississippi	South Dakota	Idaho	Oklahoma
Connecticut	Montana	Tennessee	Illinois	Rhode Island
Delaware	New Hampshire	Texas	Iowa	Utah
Indiana	New Jersey	Virginia	Kansas	Vermont
Kentucky	New Mexico	Washington	Michigan	West Virginia
Louisiana	New York	Wisconsin	Missouri	Wyoming

Source: State Department of Education Web sites.

Recommendations

Regardless of the process a particular state uses to set cut scores on its tests, there are steps all states can take to ensure quality as well as transparency for the public. States should:

- *Make the score setting process and the results more transparent and accessible*. Some states make it easy to find out how they set cut scores. Others make it a test of determination. On their department of education Web sites, states should clearly explain how cut scores are set and post data about the score setting process. While states should not identify judges by name, they should also include the range of scores the judges considered and some information about what kinds of people participated in the score setting. This information would allow the media and the public to make better judgments about state educational progress.
- *Include outside representatives on score setting panels to improve alignment and help ensure rigor.* When setting cut scores some states rely solely on the judgment of grade-level teachers and curriculum specialists. This is too insular. States should include teachers from higher grades on score setting committees to help ensure that cut scores reflect an informed understanding of the skills and knowledge students need for future success. For instance, states should tap high school math teachers to help set the cut scores for elementary and middle school math tests. They should also use representatives from higher education for the cut-score process for high school assessments.
- *Validate tests in an ongoing manner.* Policymakers should regularly revisit their standards and tests to ensure that they are aligned with state policy goals. States should also validate their state tests against other measures of student performance like the National Assessment of Educational Progress (NAEP) and map achievement criteria against real-world competencies such as the ability to read a newspaper op-ed, make sense of a graph or write a coherent essay. As state data systems become more developed and track student performance through postsecondary education, states will be able to examine how student performance at different points in elementary and secondary education lines up with postsecondary attainment.

The print and broadcast media should also take steps to ensure that the public has a full understanding of achievement data by reporting the whole story. While the media frequently reports on test scores, they rarely discuss cut scores. To help the public understand what test scores mean, the media should describe how cut scores are set and at a minimum, news stories should tell readers what a passing score of proficient actually means. Members of the media should also become familiar with how cut scores are set and benchmark state test scores against other measures like NAEP in order to give the public a full picture of the state's educational achievement.

Appendix 1. South Dakota 2005 Raw and Scaled Score Cut Points and Performance Standards

Grade 3 - State Performance Standards				
Subtest	**Below Basic**	**Basic**	**Proficient**	**Advanced**
Reading Scale Score	497 or Below	498–594	595–663	664 or Above
Reading Raw Score	0–4	5–23	24–40	41–48
Math Scale Score	502 or Below	503–589	590–643	644 or Above
Math Raw Score	0–13	14–59	60–88	89–105
Grade 4 - State Performance Standards				
Subtest	**Below Basic**	**Basic**	**Proficient**	**Advanced**
Reading Scale Score	524 or Below	525–606	607–667	668 or Above
Reading Raw Score	0–6	7–25	26–41	42–51
Math Scale Score	512 or Below	513–611	612–663	664 or Above
Math Raw Score	0–7	8–49	50–79	80–105
Grade 5–State Performance Standards				
Subtest	**Below Basic**	**Basic**	**Proficient**	**Advanced**
Reading Scale Score	543 or Below	544–620	621–677	678 or Above
Reading Raw Score	0–6	7–26	27–43	44–56
Math Scale Score	551 or Below	552–634	635–681	682 or Above
Math Raw Score	0–10	11–50	51–78	79–105
Grade 6–State Performance Standards				
Subtest	**Below Basic**	**Basic**	**Proficient**	**Advanced**
Reading Scale Score	555 or Below	556–635	636–691	692 or Above
Reading Raw Score	0–5	6–25	26–42	43–56
Math Scale Score	575 or Below	576–657	658–704	705 or Above
Math Raw Score	0–15	16–60	61–85	86–105
Grade 7–State Performance Standards				
Subtest	**Below Basic**	**Basic**	**Proficient**	**Advanced**
Reading Scale Score	564 or Below	565–650	651–708	709 or Above
Reading Raw Score	0–5	6–27	28–44	45–56
Math Scale Score	588 or Below	589–673	674–732	733 or Above
Math Raw Score	0–10	11–51	52–85	86–105
Grade 8–State Performance Standards				
Subtest	**Below Basic**	**Basic**	**Proficient**	**Advanced**
Reading Scale Score	573 or Below	574–667	668–727	728 or Above
Reading Raw Score	0–5	6–26	27–41	42–49
Math Scale Score	586 or Below	587–685	686–734	735 or Above
Math Raw Score	0–7	8–51	52–80	81–105
Grade 11–State Performance Standards				
Subtest	**Below Basic**	**Basic**	**Proficient**	**Advanced**
Reading Scale Score	624 or Below	625–718	719–779	780 or Above
Reading Raw Score	0–5	6–24	25–35	36–40
Math Scale Score	577 or Below	578–702	703–758	759 or Above
Math Raw Score	0–3	4–46	47–80	81–105

Source: South Dakota Department of Education.

Endnotes

[1] See Catherine Horn, Miguel Ramos, Irwin Blumer, and George Madaus, *Results May Vary* (Boston, MA: National Board on Educational Testing and Public Policy, Peter S. and Carolyn A. Lynch School of Education, 2000).

[2] There is a great deal of debate about the quality of state standards. See, for instance, Making Standards Matter (Washington, DC: American Federation of Teachers, 2001); The State of Math Standards, 2005 (Washington, DC: Thomas B. Fordham Foundation, 2005); The State of English Standards 2005 ((Washington, DC: Thomas B. Fordham Foundation, 2005). See also, Bella Rosenberg, *What's Proficient: The No Child Left Behind Act and the Many Meanings of Proficiency* (Washington, DC: American Federation of Teachers, 2004).

[3] There are less common methods. See, for instance, Samuel A. Livingston and Michael J. Zieky, *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests* (New Jersey: Educational Testing Service, 1982).

[4] Livingston, *op cit*.

[5] *Ibid*.

[6] Wisconsin Department of Public Instruction, "Bookmark Standards Setting," http://dpi.wi.gov/oea/ctbbkmrk03.html. Retrieved June 30, 2006.

[7] Livingston, *op cit*.

[8] Education Sector Research Assistant Margaret Price surveyed state education departments from May to July, 2006.