Purpose of the test – To evaluate student proficiency

The important point I wish the board members to understand is what exactly is the difference between a test like NECAP, designed to rank schools and students, and a test designed to evaluate student proficiency. The short version: when you design a test like NECAP, test designers ensure that a certain number of students will flunk. What's more, for the purposes of the test designers, that's a good thing.

- The NECAP tests were designed specifically to evaluate student proficiency.
- The NECAP tests were designed to meet the assessment and accountability requirements of No Child Left Behind. Although a primary use of assessment results under NCLB was school and district accountability, the accountability model has shifted from ranking schools and students. In the standards-based era of NCLB, contrary to "ensuring that a certain number of students will flunk" the measure of school accountability was the percentage of its students demonstrating performance at the Proficient level or higher and the goal was 100% of students Proficient.
- The results of the Grade 11 NECAP Reading test bear this out. On the most recent Fall 2012 test, 79% of students performed at the Proficient level or higher (up from 76% the previous two years) and 92% of students met the student graduation requirement of Partially Proficient.
    - One-third of grade 11 students (33 percent) scored at the highest achievement level on the Grade 11 Reading test.

Student Performance on Individual Items

In other words, very few of the questions are correctly answered by all students. In Appendix F of the 2011-12 manual, you can see some item-level analyses. There, one can read that, of the 22 test questions analyzed, there are no questions on the 11th grade math test correctly answered by more than 80% of students, and only nine out of 22 were correctly answered by more than half the students.

Put another way, if all the students in a grade answered all the questions properly, the NECAP designers would consider that test to be flawed and redesign it so that doesn't happen. Much of the technical manual, especially chapters 5 and 6 (and most of the appendices), are devoted to demonstrating that the NECAP test is not flawed in this way. Again, the NECAP test is specifically designed to flunk a substantial proportion of students who take it, though this is admittedly a crude way to put it.

- The item statistics cited in Appendix F of the 2011-2012 Technical Report apply to only the 22 1-point or 2-point short-answer and 4-point constructed-response items included on the test. These items account for 40 of the 64 points on the Grade 11 Mathematics test. Historically, these items which require students to produce a response are more difficult than the multiple-choice items which require students to select a correct response.
- Item statistics for the multiple-choice items on the Grade 11 Mathematics test are presented in Appendix E of the same Technical Report. Across those 24 items, 10 were answered correctly by more than half the students and two were answered correctly by at least 80% of the students.
- Once again, however, results from the Grade 11 Reading test demonstrate that the item statistics cited by the author are more a reflection of student performance in mathematics than intentional test design. On the reading test, there are 28 multiple-choice items. Across those items, 27 of 28 were answered correctly by more than half the students, with 49% answering the remaining item correctly. Additionally, at least 80% of students answered eight of the reading items correctly with 90% of students answering one of the items correctly.

Impact of Measurement Error
Furthermore, like any other measurement, a test score has an inherent error. For any individual student, a teacher can have little confidence that a student who scored an 80 didn't deserve an 84 because of a bad day, a careless mistake, or, worse, someone else's error: a misunderstood instruction, an incomplete erasure, or a grading mistake. Of course, any errors could also move the score in the other direction.

- Yes, measurement error is present in any test score. On the grade 11 NECAP mathematics test the standard error or measurement near the Partially Proficient cut score of 1134 required for graduation is approximately 2 scaled score points. That standard error is taken into account in two important ways with regard to the student graduation requirement.
  - The Board has set the minimum score on the NECAP tests for student graduation at the *Partially Proficient* level. This is well below the *Proficient* level that is the goal for all students and the requirement for school and district accountability. Note that Proficient is the level of performance met by 79% of the grade 11 students on the Reading test. There is less than a 1 in 1,000 chance that a student who is actually performing at the Proficient level or higher will score below the Partially Proficient cut due to measurement error present in the test score.

  - Of course, there is a greater, but still small, chance of "false negatives" among students whose performance is very close to the Partially Proficient cut score. Among those students there is a 2%-3% chance of a student who is actually Partially Proficient scoring below the graduation requirement on a **single** administration of the test. The chance of a false negative due to measurement error declines dramatically with every additional opportunity to demonstrate proficiency. After 3 opportunities to take the test the likelihood of a false negative due to measurement error is well below 1 percent. That is the primary reason why no graduation decisions are based on a single administration of the NECAP test. In accordance with professional standards and established practices, students are provided multiple opportunities to meet the state assessment graduation requirement through two opportunities to retake the NECAP test or by providing evidence of proficiency from other approved, external assessments. In addition, the regulations allow for waivers to the state assessment requirement to be provided in those rare cases in which there is clear evidence that a standardized assessment is not a valid measure of student performance.

  - As the author correctly points out, "any errors could also move the score in the other direction." On the NECAP tests, the rate of false positives at the Partially Proficient level is approximately 2%-4%, consistent with the rate of false negatives. In the case of high-stakes graduation decisions, established practice reflects that false negatives have more serious consequences (i.e., denial of a diploma) than false positives.
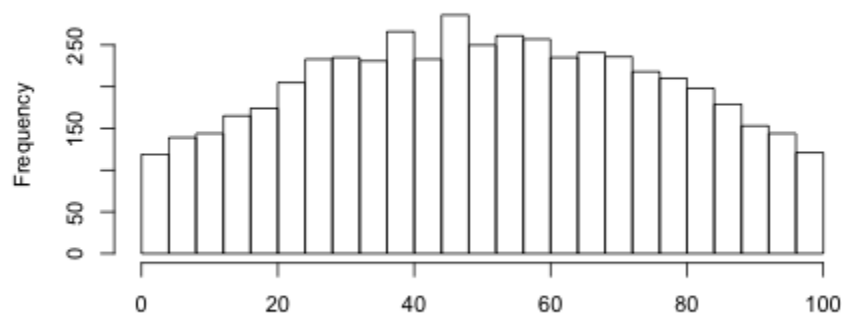
Distribution of Student Scores

The author presents two figures as examples of distributions of student scores on a test. The first is presented as the type of skewed distribution one might hope to see in a test designed to measure student proficiency. The second is a distribution of scores that the author claims is the goal of the NECAP tests.

If the goal is to see which of the students in the class have properly understood the material, this is a useful result.
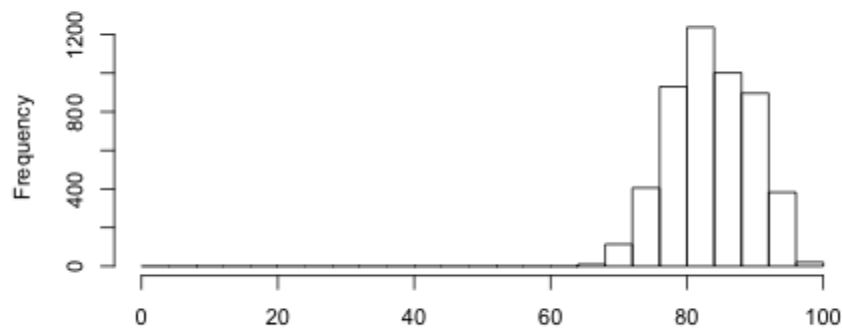

Instead, they try to design tests so the distribution of scores looks more like the one here:


The two figures on the following page present the distribution of student scores from the Grade 11 reading and mathematics tests. Comparing those results to the figures above, although one is for a test on which students performed well (reading) and one is for a test on which student performance is poor (mathematics) it is clear that both distributions are skewed in a way that reflects student proficiency (similar to the type of "desirable" distribution in the first figure above) rather than attempting to force a normal distribution centered in the middle of the score scale.

Fall 2012 Grade 11 NECAP Mathematics Test
Distribution of Student Scores

Mathematics Ach. Level: SBP, PP, Prof, PWD
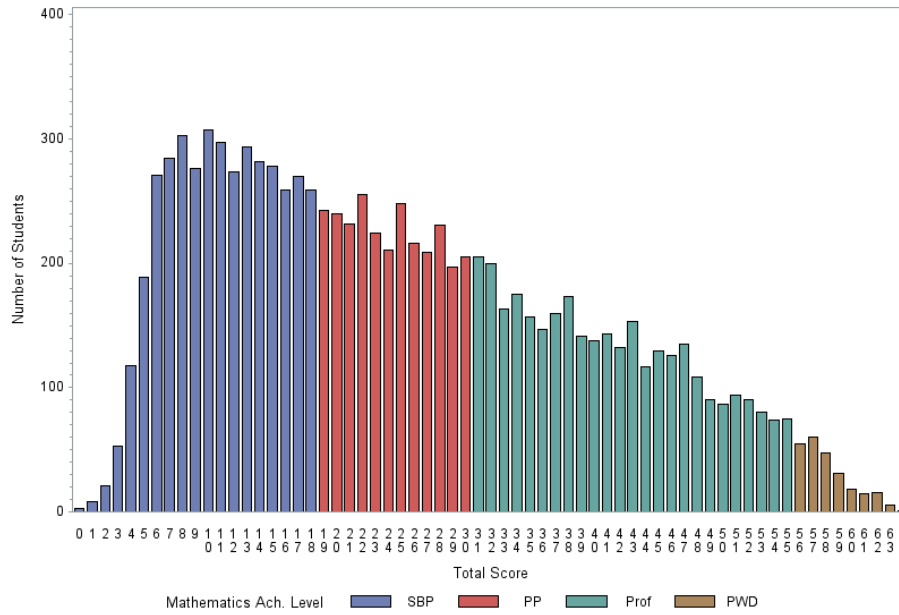
Fall 2012 Grade 11 NECAP Reading Test
Distribution of Student Scores

Reading Ach. Level: SBP, PP, Prof, PWD
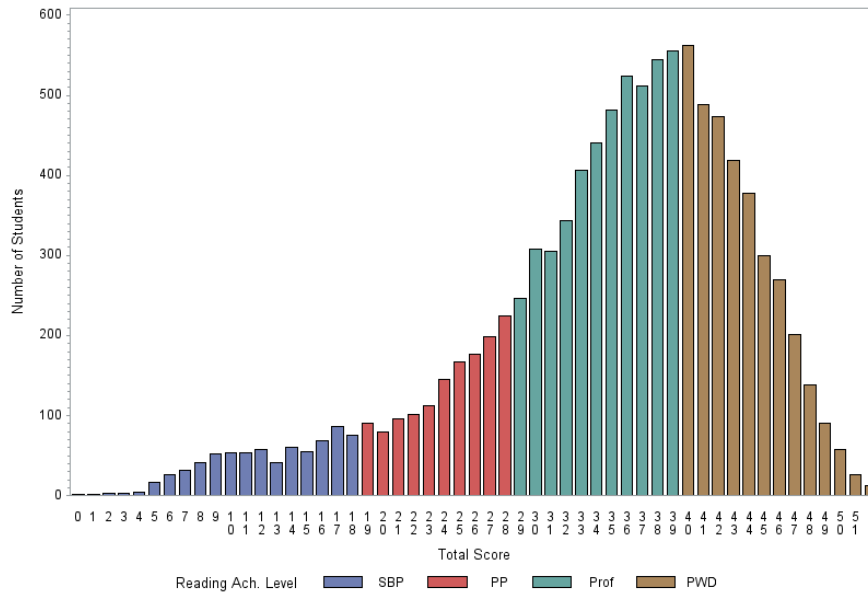
Content on the Grade 11 NECAP Mathematics Test

11th Grade Math *Before leaving the subject of students flunking the NECAP tests, it's worth taking a moment to consider the 11th grade math test specifically. …*

However, it is worth noting that the tests occur almost two years before a student's graduation, and that math education proceeds in a fundamentally different way than reading. That is, anyone who can read at all can make a stab at reading material beyond their grade level, but you can't solve a quadratic equation halfway.

Rather than providing a measure of student competence on graduation, the test might instead be providing a measurement of the pace of math education in the final two years of high school. The NECAP test designers would doubtless be able to design questions or testing protocols to differentiate between a good student who hasn't hit the material yet, or a poor student who shouldn't graduate, but they were not tasked with doing that, and so did not.

- There is no requirement or expectation for students to "make a stab" at "material beyond their grade level" on the Grade 11 NECAP tests.
- The Grade 11 NECAP tests, administered in October of the eleventh grades, are designed to measure student achievement of the Grade 9-10 content standards. In mathematics, those standards address topics covered primarily in Algebra I and Geometry courses.
- The state assessment portion of the graduation requirements in both Reading and Mathematics specifically are limited to student performance through grade 10.
- The other two school-based dimensions of student graduations requirements (coursework and performance-based portfolios or exhibitions) focus more on performance over all four years of high school.

Performance of Students Not Meeting the Graduation Requirement

On the following page are selected sample items from the Fall 2012 Grade 11 NECAP mathematics test which show the level of mathematics being assessed and the performance of students not meeting the graduation requirement who answered those items correctly.

Question 13 - 6%

⑬ A real-estate agent receives a 3.5% commission on the sale of a house that costs $200,000. What is the amount, in dollars, of the commission? [commission = sale price × rate]

Question 18 – 11%

⑱ This table shows the number of available lockers at a gym.

Locker Numbers	Number of Available Lockers
100–199	40
200–299	35
300–399	25
400–499	20

One of the available lockers is randomly assigned to a new member at the gym. What is the probability that the locker number starts with a 3 or a 4?

Question 5 – 12%

⑤ The center of circle P is located at (5, –2).

- $\overline{ST}$ is a diameter of circle P.
- Point T is located at (–3, 3).

What are the coordinates of point S?

A. (1, 0.5)
B. (4, 2.5)
C. (8, –5)
D. (13, –7)

Question 2 – 20%

❷ The current value of an investment is 125% of its initial value. The increase in value was $10 million. What was the initial value of the investment?

A. $ 25 million
B. $ 40 million
C. $100 million
D. $115 million