



Introduction to Measurement

AB 250 Panel Presentation
April 17, 2012

Eric Zilbert, Administrator
Assessment Development and Administration Division



TOM TORLAKSON
State Superintendent
of Public Instruction

Presentation Overview

- Test purpose as the basis for creating measures
- Considerations for assessment program development
- Assessment development process
- Reliability and Validity of Assessments
- Test Fairness
- Test Formats



TOM TORLAKSON
State Superintendent
of Public Instruction

Bases for Creating Measures

- Purpose for which test results will be used is the key.
- Purposes include: School accountability, pupil achievement, diagnosis, certification
- More purposes = more testing
- Using a test for purposes other than that for which it is designed invalidates the test results.
- The domain (content) to be measured must be clearly defined.
- A measurement model must be specified that explains how the test will sample from the domain to be assessed. The measurement model and content inform the design of the test blueprint and test specifications.



TOM TORLAKSON
State Superintendent
of Public Instruction

Considerations for Different Test Purposes

For School and District Accountability

- Forms need to be changed from year to year to protect integrity of the tests.
- Require very accurate scaling and equating of tests.
- Test security is essential
- Independently scored
- Criterion referenced
- Administration conditions need to be as standardized as possible to insure fair comparisons.
- Matrix testing can be used to better assess coverage of the curriculum without extending testing time
- Generally large scale assessments need to keep scoring costs low, hence the dependence on multiple choice and other dichotomous items.



TOM TORLAKSON
State Superintendent
of Public Instruction

Considerations for Different Test Purposes

For Measuring Student Achievement and Growth

- Can be norm or criterion referenced
- Accuracy is primarily a function of test length
- Students may take different items, tests must be carefully scaled and equated for fair comparisons.
- Growth measurement is best achieved where standards are linked across grades in “learning progressions.”
- Measurement of growth may be achieved with vertically linked assessments or computer adaptive testing.
- Most accurate when a limited portion of the domain is measured (e.g., reading comprehension).
- Security may or may not be a concern depending on consequences of test results (school accountability, placement, awards, student grades, teacher evaluation etc.)

Considerations for Different Test Purposes

For Diagnostic or Formative Purposes

- Test items must be tightly linked to the curriculum, much narrower scope than a year end (summative) assessment.
- Require more testing time than achievement tests for accountability purposes.
- Equating is not as serious an issue, as questions address specific skills or knowledge that a student either does or does not possess (e.g., ability to do long division with two to four digit numbers).
- Number correct is often sufficient information to make decisions about student placement.
- Security generally not as big an issue. May be scored by teachers, same forms may be used over and over.
- Provide directly actionable information and are closely tied to instruction (Johnny needs to learn how to borrow when subtracting).



TOM TORLAKSON
State Superintendent
of Public Instruction



TOM TORLAKSON
State Superintendent
of Public Instruction

Considerations for Different Test Purposes

For Certification

- e.g., graduation examinations like the CAHSEE
- Minimization of error at the cut score a major concern
- Usually incorporate constructed response items involving complicated stimulus designed to reflect skills and knowledge needed for success.
- Cost of test development and scoring generally borne by the candidates.



TOM TORLAKSON
State Superintendent
of Public Instruction

Considerations in Test Program Development

- Purpose
- Resources
- Psychometrics
- Interface
(including accommodations)
- Item designs
- Test designs
- Test distribution
- Item exposure
- Item and test security
- Examiner and proctor training
- Scoring
- Reporting



TOM TORLAKSON
State Superintendent
of Public Instruction

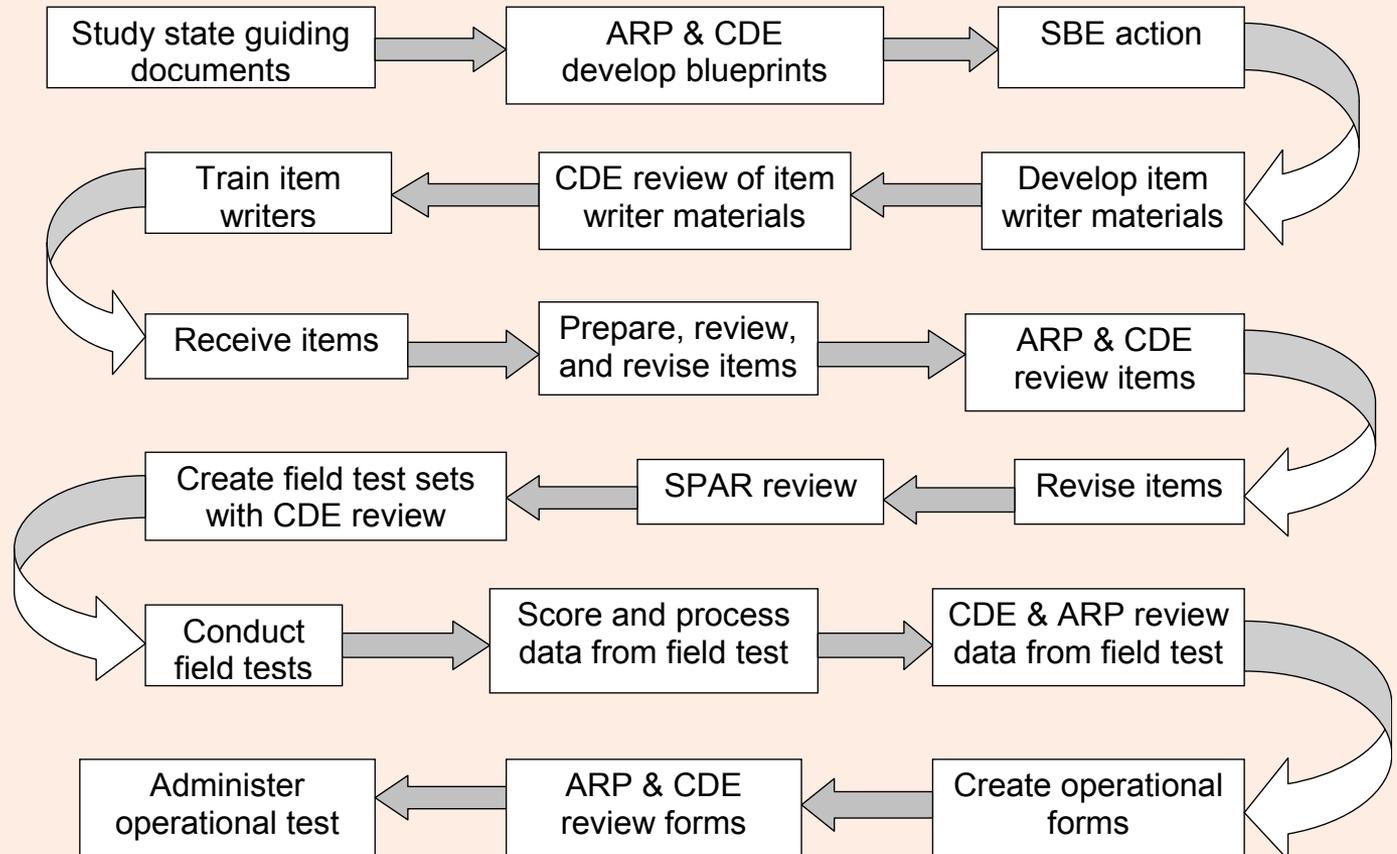
Item Types

- Selected Response
 - Correct/Incorrect scoring, e.g., T/F, multiple choice, graphic plot, bubble a numeric response
 - Can be quickly and easily scored
- Short Constructed Response
 - Include short answer and fill in the blank item types
 - May have several score points
- Extended Constructed Response
 - Answer with explanation, essay, and performance based items
 - Historically more time consuming and costly to score
 - Recent developments in automated scoring systems have the potential to reduce time and cost required to score



TOM TORLAKSON
State Superintendent
of Public Instruction

Test Development Process



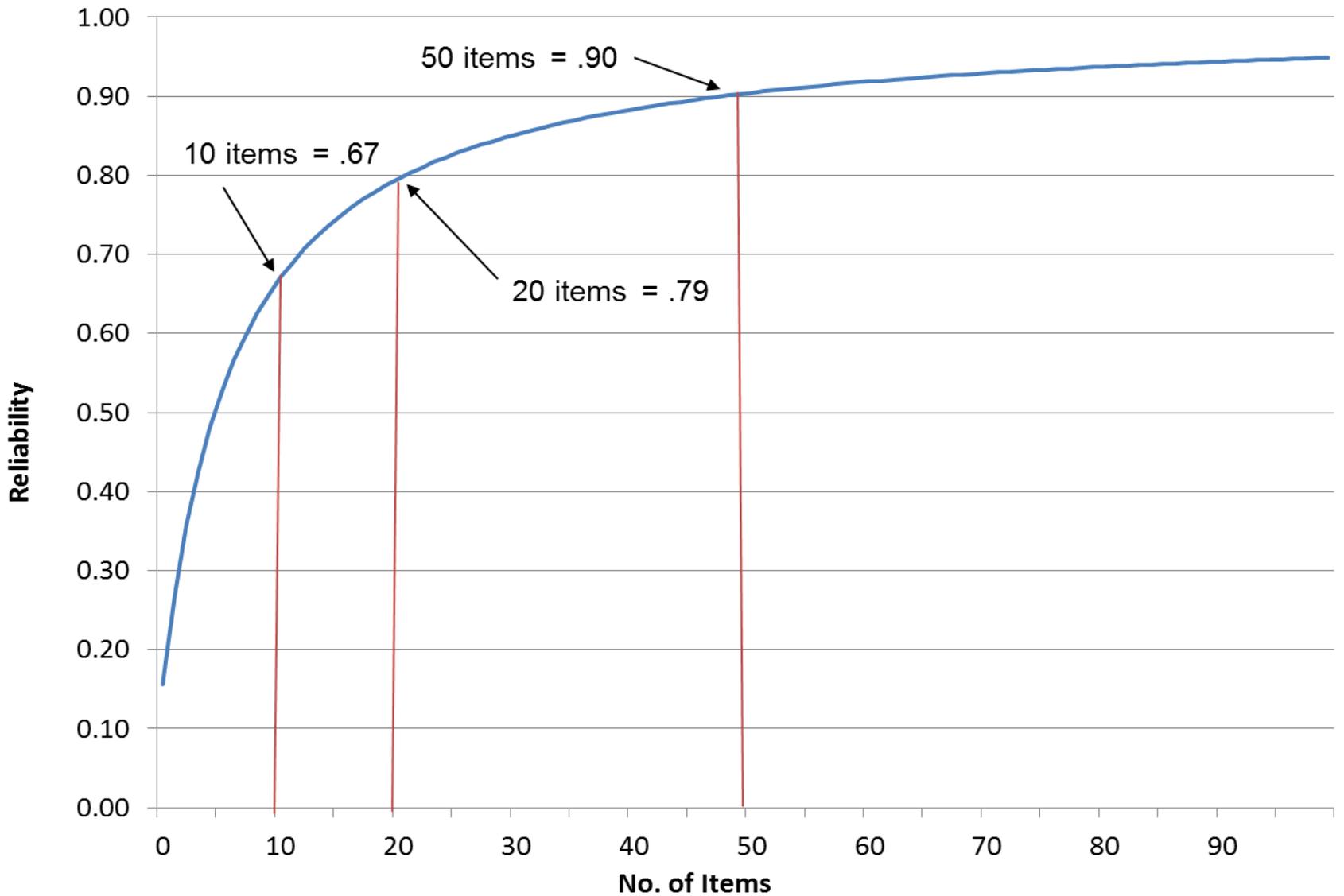


TOM TORLAKSON
State Superintendent
of Public Instruction

Reliability and Validity

- **Reliability**
 - How consistently does the test measure the underlying construct
 - Can be determined statistically (e.g., Cronbachs Alpha, Test-Retest)
- **Validity**
 - Evidence that the test is appropriate for the intended use of the results
 - It is the interpretation of the test score and how it is used that is validated, not the test
 - Mostly based on human judgment

Reliability and Test Length





TOM TORLAKSON
State Superintendent
of Public Instruction

Types of Validity Evidence

- **Content review by experts (face validity)**
 - Item review by Assessment Review Panels (ARP), Independent Alignment Reviews, tec.
- **Relation to other criteria (concurrent or predictive validity)**
 - Evidence based on related measures. E.g., CAHSEE ELA vs. CST Grade 10 ELA
 - Driving test vs. written drivers test.
- **Evidence based on internal consistency**
 - Statistical measures of how well the items relate to the construct being measured (e.g., point biserial correlation)
- **Evidence based on consequences of testing (consequential validity)**
 - Do the test results bring about the desired consequences? Examples: Are high scoring students more likely to succeed in college or in the workplace?
Are schools improving?



TOM TORLAKSON
State Superintendent
of Public Instruction

Validity Evidence Through Alignment Reviews

- How much content is covered by the assessment?
- Is this content sufficiently similar to the expectations of the standards?
- Are students asked to demonstrate this knowledge at the same level of rigor as expected in the content standards?



TOM TORLAKSON
State Superintendent
of Public Instruction

Alignment Terminology (Webb)

Categorical concurrence – the proportion of overlap between the content stated in the standards document and that assessed by items on the test.

Depth-of-Knowledge – (DOK) measures the type of cognitive processing required by items and content standards.

Range-of-knowledge – indicates the number of content objectives assessed by items.

Balance-of-knowledge – evaluates the degree standards are equitably represented for each content strand.



TOM TORLAKSON
State Superintendent
of Public Instruction

Test Fairness

- Opportunity to learn - It is generally agreed that there is little to be gained from testing students on information that they have not been asked to learn.
- Tests should not present barriers to a student's ability to demonstrate what they know.
- All students should have an equal opportunity to demonstrate what they know and have a chance to respond to items of the same range of difficulty as others to whom they will be compared.
- Students with disabilities and English learner students may require special accommodations to ensure that they can meaningfully participate in the assessment.
- Variations and accommodations provided to students should not provide an advantage over students that do not get to use them. Goal is to level the playing field (e.g., large print for visually impaired student.)



TOM TORLAKSON
State Superintendent
of Public Instruction

Checks for Test Fairness

- Bias and Sensitivity Reviews – Panelists review items for potential bias related to age, gender, race, ethnicity, English learner status and socio-economic status.
- Differential Item Functioning (DIF) Analysis – Statistical test to determine if items function differently for different groups.
- Universal Design Review – Review by experts to assure principles of universal design have been applied to test items and test forms.



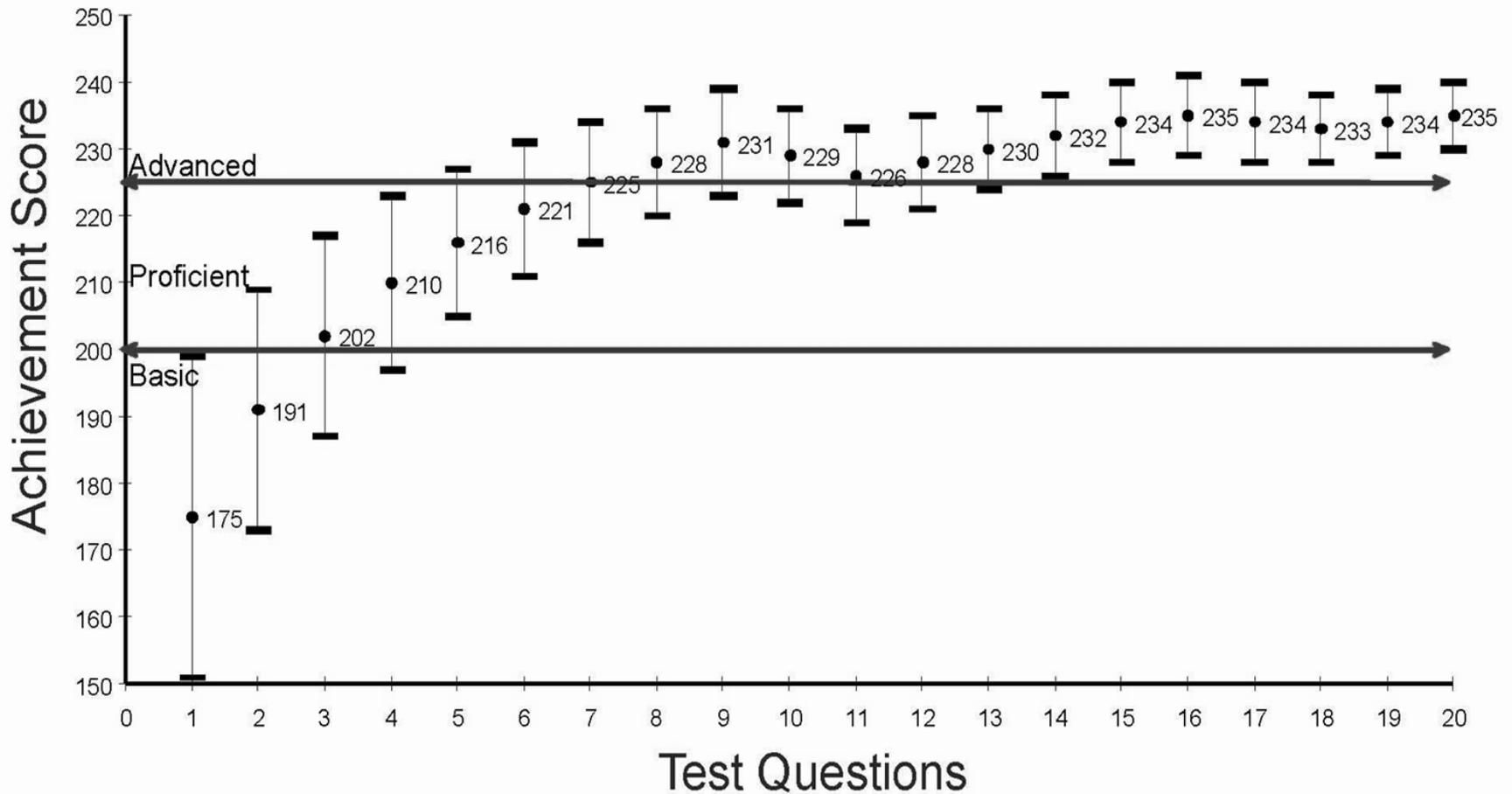
TOM TORLAKSON
State Superintendent
of Public Instruction

Test Formats

- Paper and Pencil
 - Most common type of test
 - Used for classroom and statewide assessments
- Computer Based (CBT)
 - Fixed form is used, but administered using a computer
 - Can allow wider variety of stimuli and item types
- Computer Adaptive (CAT)
 - Student receive harder or easier questions depending on how they perform as the test proceeds
 - Can provide greater accuracy with fewer questions
- Performance Assessments
 - Require students to produce a product using a variety of stimuli or resources

Computer Adaptive Testing

20 Questions



Test Information Functions CAT and Paper and Pencil Tests

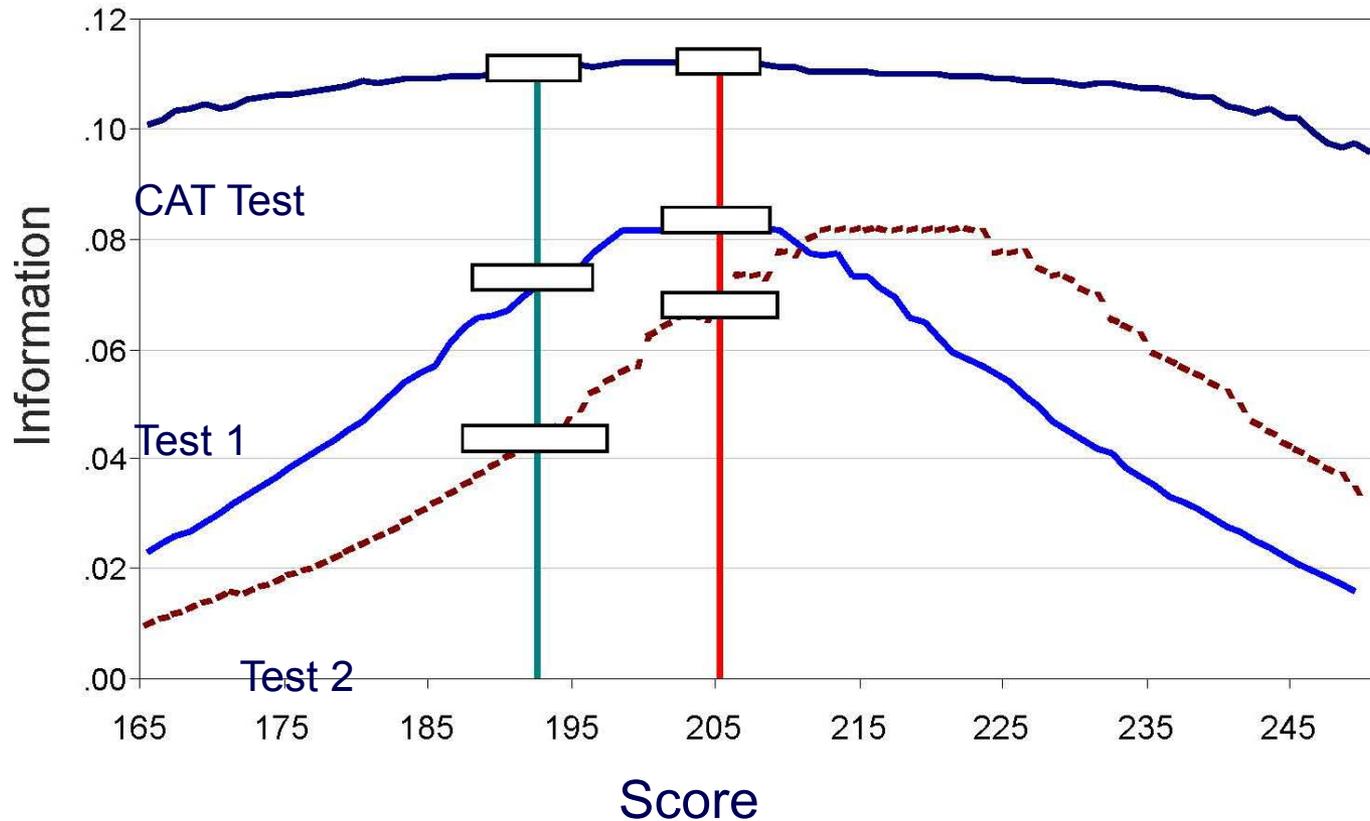
Students' Mean = 211.7

s.d. = 11.11

Proficiency = 205

Basic = 192

Test Information Functions for Grade 4 Mathematics





TOM TORLAKSON
State Superintendent
of Public Instruction

Questions?