Overcoming the Limitations of State Testing

How should we use the results from NJASK and GEPA?

NJSBA Convention 2007, Atlantic City, NJ

Dr. Patrick Michel, Haddon Heights SD

Dr. Christopher Tienken, Monroe Twp. SD

Mr. Thomas Tramaglini, Freehold Boro SD

Objectives

- By the end of the presentation you will:
  - Gain a general understanding of the technical and practical limitations of the NJ state test results
  - Be able to identify alternatives to using state test results as the only factor to make judgments about students and programs
  - Be able to design a district assessment system to better meet the needs of students and the requirements of NJQSAC

High Stakes

- High stakes decisions require the use of quality data to guide decision-making.

- Given the data from states test are used to make important decisions about students, schools, and programs, it is reasonable to expect those data to be of high quality.

High Expectations for State Test Results

- As education leaders and board members we expect:

1. The reported individual and group scores on the tests to be precise.
2. An accurate measurement and appraisal of student and program performance.
3. The test questions to provide an accurate representation of the NJCCCS

Limitations

- **All** high stakes state tests have limitations that restrict the usability of the results

- NJ's Include:

  1. Lack of test score precision (error)
  2. Low reliability of the content-cluster scores
  3. Questionable content validity (not enough questions to measure NJCCCS)

Problem: District Use of Test Results

- Statewide random sample

- Over 95% of responding districts used the results as one of multiple factors to make high stakes decisions about students and programs.

- Over 50% of responding districts used the results as the only or deciding factor to make high stakes decisions about students and programs.

Limitation 1: Test Score Precision

- Educators expect the reported score to be the *true* score.

- This is especially important at the cut-points between Partially Proficient and Proficient (i.e. 200) and Proficient and Advanced (i.e. 250)

Test Score Precision

- Precision is important because of the ways educators use the scores. For example:

1. BSI & Title I placements
2. Course access & selections in HS
3. Recommendations for other academic and co-curricular programs
4. After school academic programs

What is the True Score?

STUDENT SCORE ON
NJASK 4
LANGUAGE ARTS

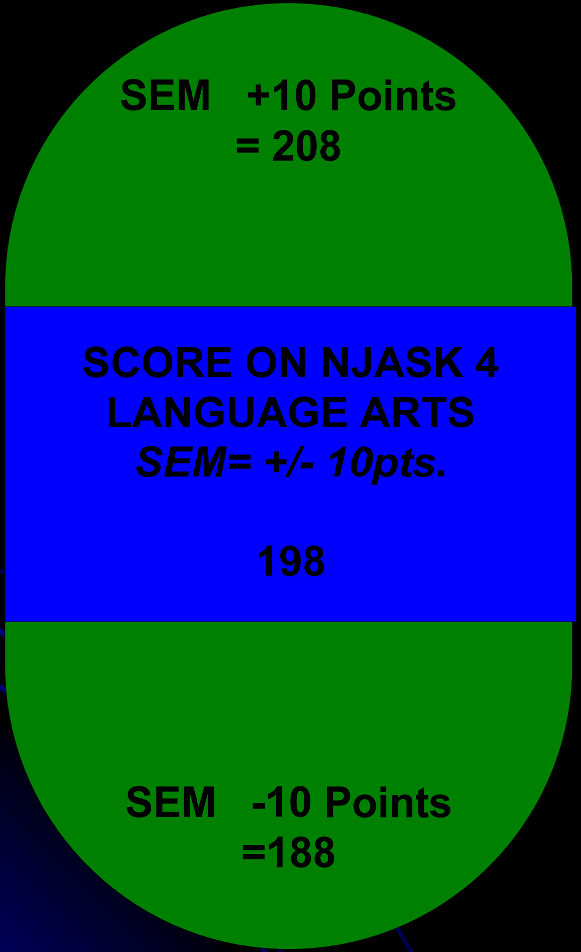
198

Scenario: Student scores a 198 on the LA portion. What do you do?

First you need to ask, *“Is this the true score?”*

Probably not, due to *measurement error*, AKA Standard Error of Measurement (SEM).

Standard Error of Measurement



SEM +10 Points
= 208

SCORE ON NJASK 4
LANGUAGE ARTS
SEM= +/- 10pts.

198

SEM -10 Points
=188

SEM is the difference between the student’s reported score and the “true” score. Think of it as the margin of error in a poll (e.g. + or – 3 pts.)

All tests have a degree of error.

Some have more than others. One would want a small amount of error when making high stakes decisions about students.

In this case, a large degree of error prevents us from making an absolute decision about this student.

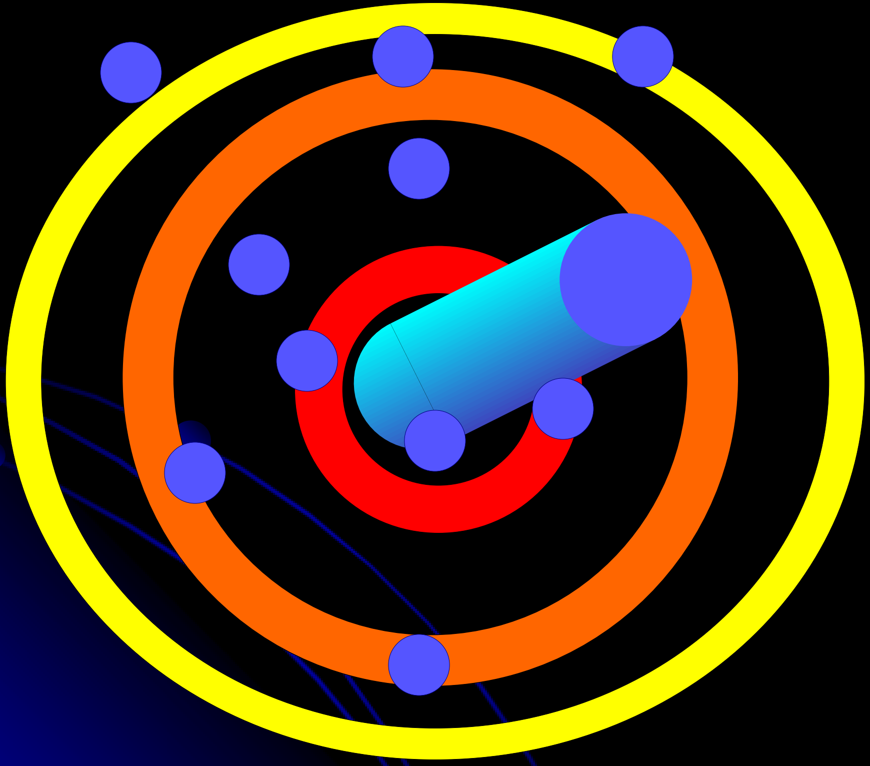
SEM and Accuracy

- One would expect that testing instruments facilitate the NJDOE & district leaders in efforts to categorize student performance accurately (i.e.):

1. Partially Proficient
2. Proficient
3. Advanced Proficient

As you can see, Standard Error of Measurement can influence accuracy of those categorizations -

SEM and Accuracy



SEM influences the accuracy of achievement categorizations

***For example*: Up to 13,000 students in math in grades 3-8 and 11 can be mis-categorized as partially proficient on state tests.**

The achievement levels of up to 91,000 students statewide may be mis-categorized in math. Some of them are your students.

Measurement Error: So What?

1. **The amount of error present in state test results for individual students should cause us to be cautious about the ways we use results to judge students.** ***The score you see may not be the "true" score.***

Limitation 2: Accuracy and Reliability

- Would your students get similar scores if they took a test on several different occasions or two different forms of the same test?
  - These questions have to do with the *consistency* with which tests measure students' achievement. The generic name for consistency is reliability.

- One wouldn't trust bathroom scales if the reading fluctuated according to the temperature or humidity or if the scales had a loose spring. Similarly, we can't trust scores from tests unless we know the consistency with which they measure. Only to the extent that scores are reliable can they be useful and fair to students. Jacobs, L. (1991)

Reliability: So What?

2. The low levels of score reliability present in state tests should cause us to be cautious about making high stakes decisions about the effectiveness of curriculum and instruction or student achievement.

Experts recommend .90-.95 reliability estimate to make high stakes decisions about individual students.

GEPA Reliability (NJDOE, 2005, p.53)

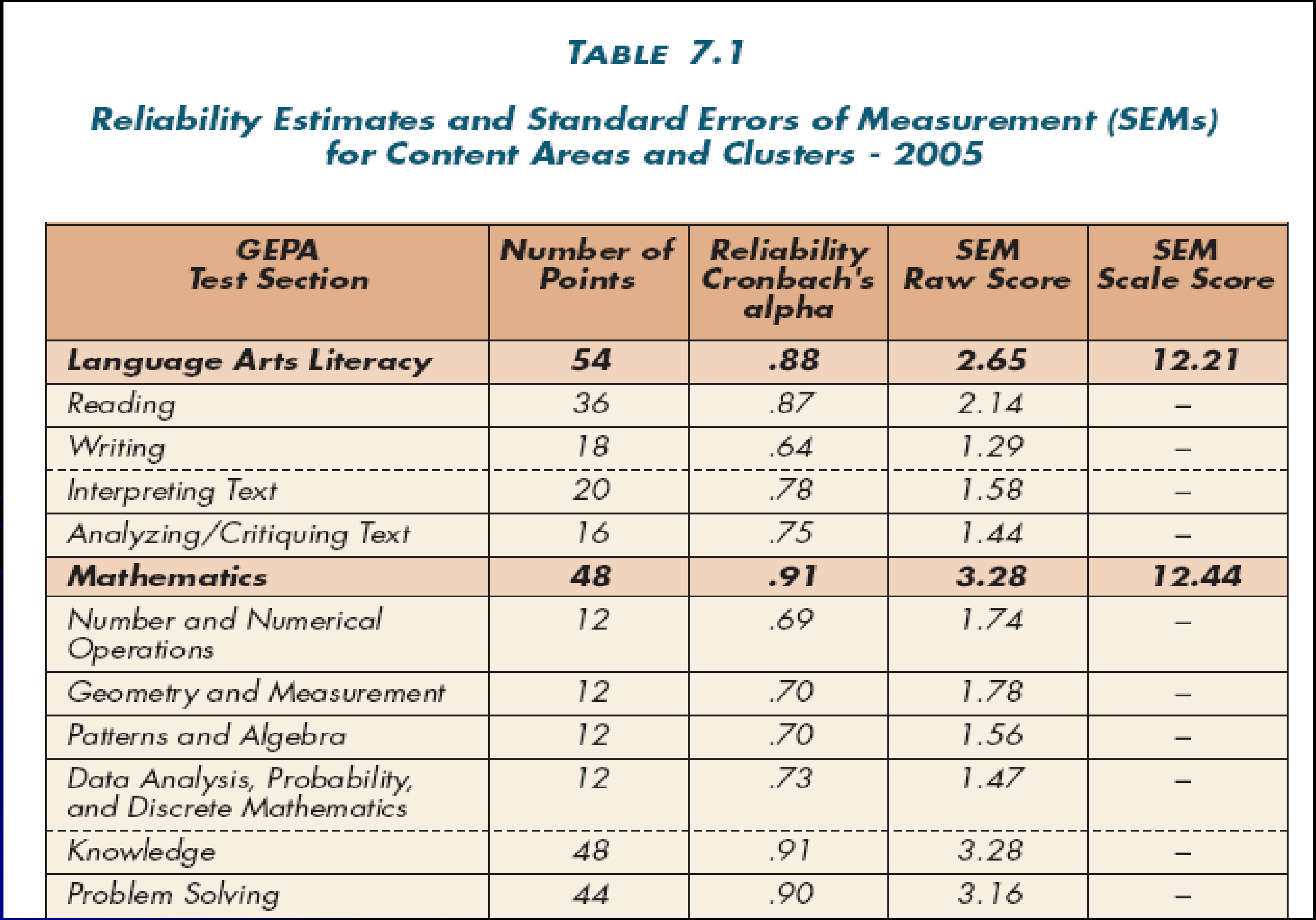


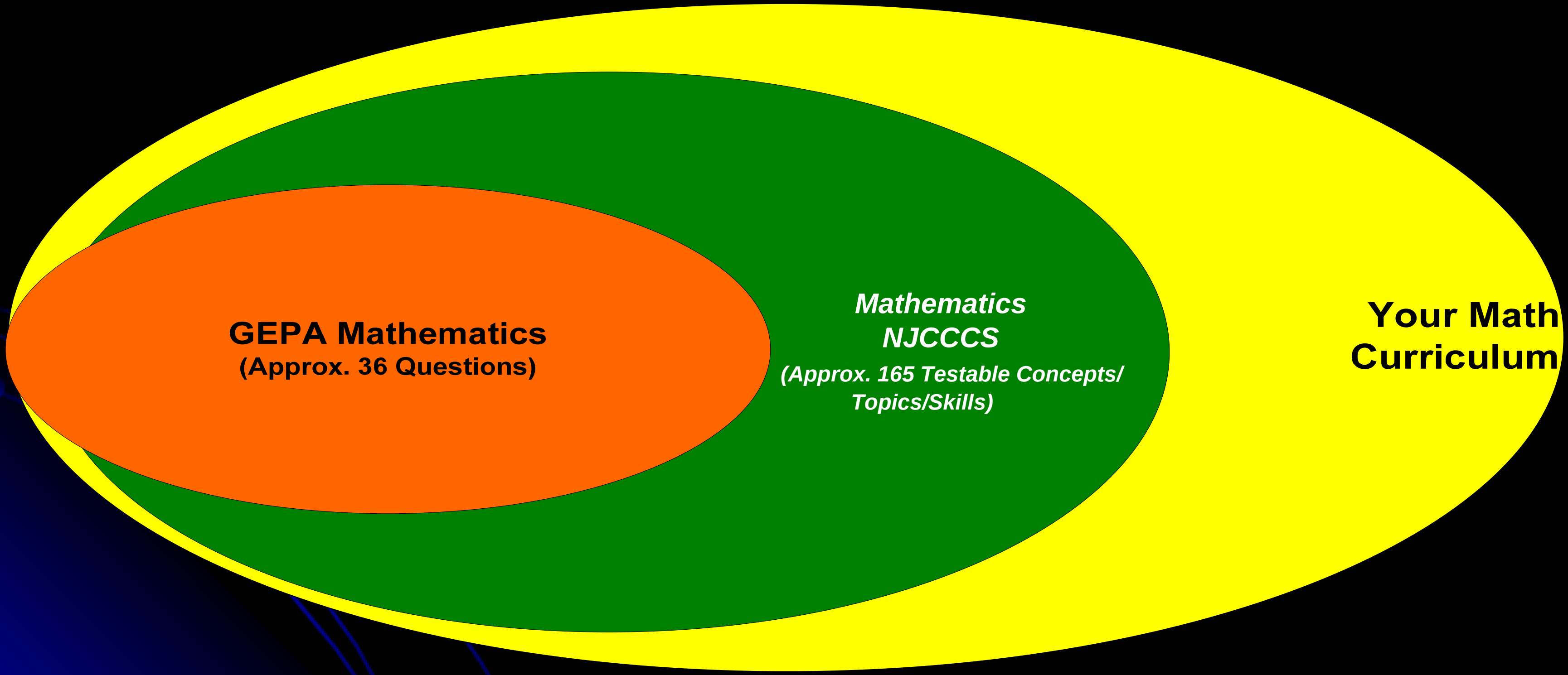
TABLE 7.1

Reliability Estimates and Standard Errors of Measurement (SEMs)
for Content Areas and Clusters - 2005

GEPA Test Section	Number of Points	Reliability Cronbach's alpha	SEM Raw Score	SEM Scale Score
Language Arts Literacy	**54**	**.88**	**2.65**	**12.21**
Reading	36	.87	2.14	–
Writing	18	.64	1.29	–
Interpreting Text	20	.78	1.58	–
Analyzing/Critiquing Text	16	.75	1.44	–
Mathematics	**48**	**.91**	**3.28**	**12.44**
Number and Numerical Operations	12	.69	1.74	–
Geometry and Measurement	12	.70	1.78	–
Patterns and Algebra	12	.70	1.56	–
Data Analysis, Probability, and Discrete Mathematics	12	.73	1.47	–
Knowledge	48	.91	3.28	–
Problem Solving	44	.90	3.16	–

Limitation 3: Content Coverage

Educators expect the tests to measure what they claim to measure (Content Validity) –



GEPA Content Coverage?

(NJDOE, 2005, p.53)

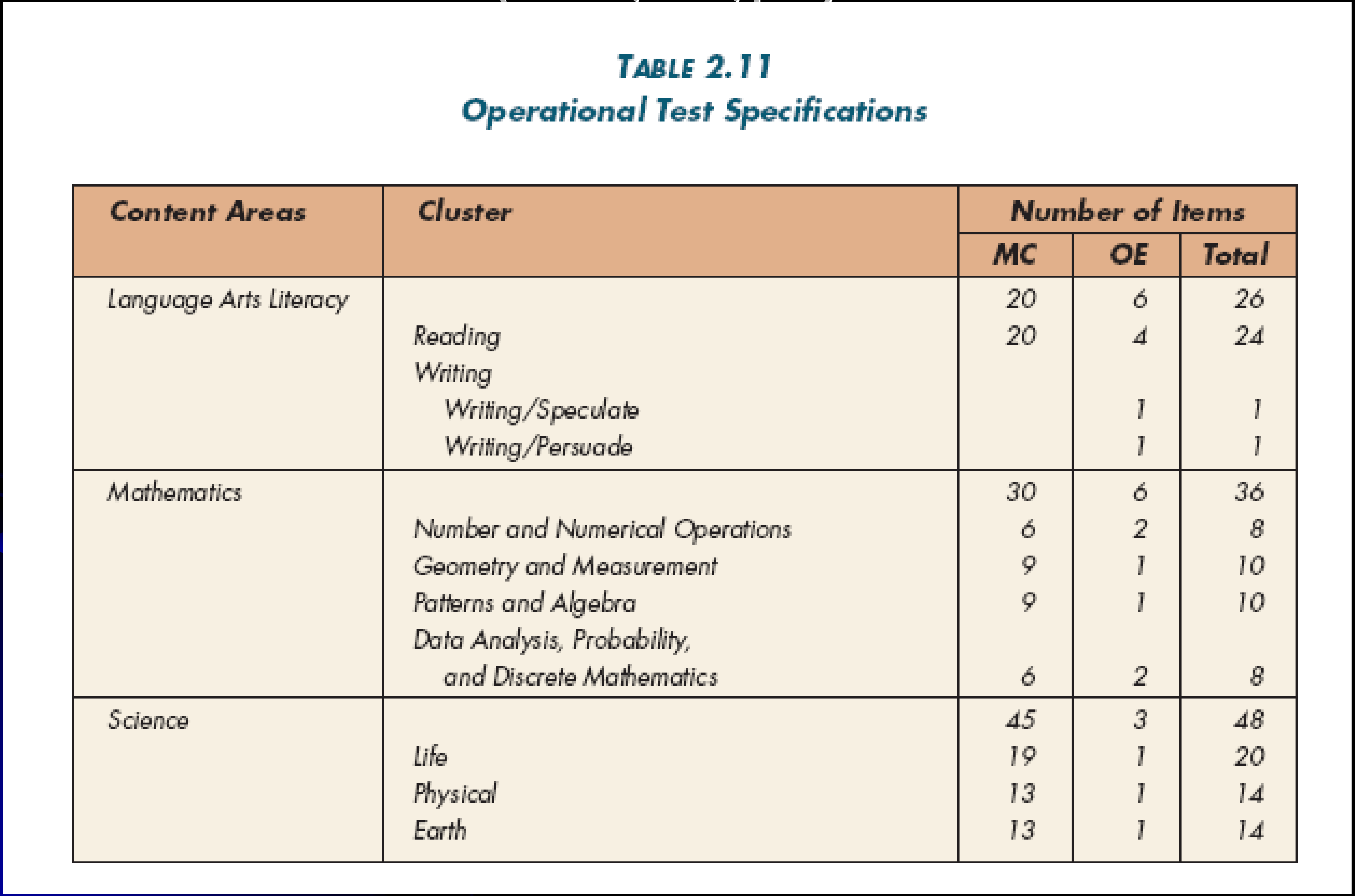


TABLE 2.11
Operational Test Specifications

Content Areas	Cluster	Number of Items		
		MC	OE	Total
Language Arts Literacy		20	6	26
	Reading	20	4	24
	Writing			
	Writing/Speculate		1	1
	Writing/Persuade		1	1
Mathematics		30	6	36
	Number and Numerical Operations	6	2	8
	Geometry and Measurement	9	1	10
	Patterns and Algebra	9	1	10
	Data Analysis, Probability, and Discrete Mathematics	6	2	8
Science		45	3	48
	Life	19	1	20
	Physical	13	1	14
	Earth	13	1	14

Content Coverage: So What?

3. The content coverage of the state tests should cause us to be cautious about the way we use results to make high stakes decisions about how much students know and can do and about the effectiveness of programs.

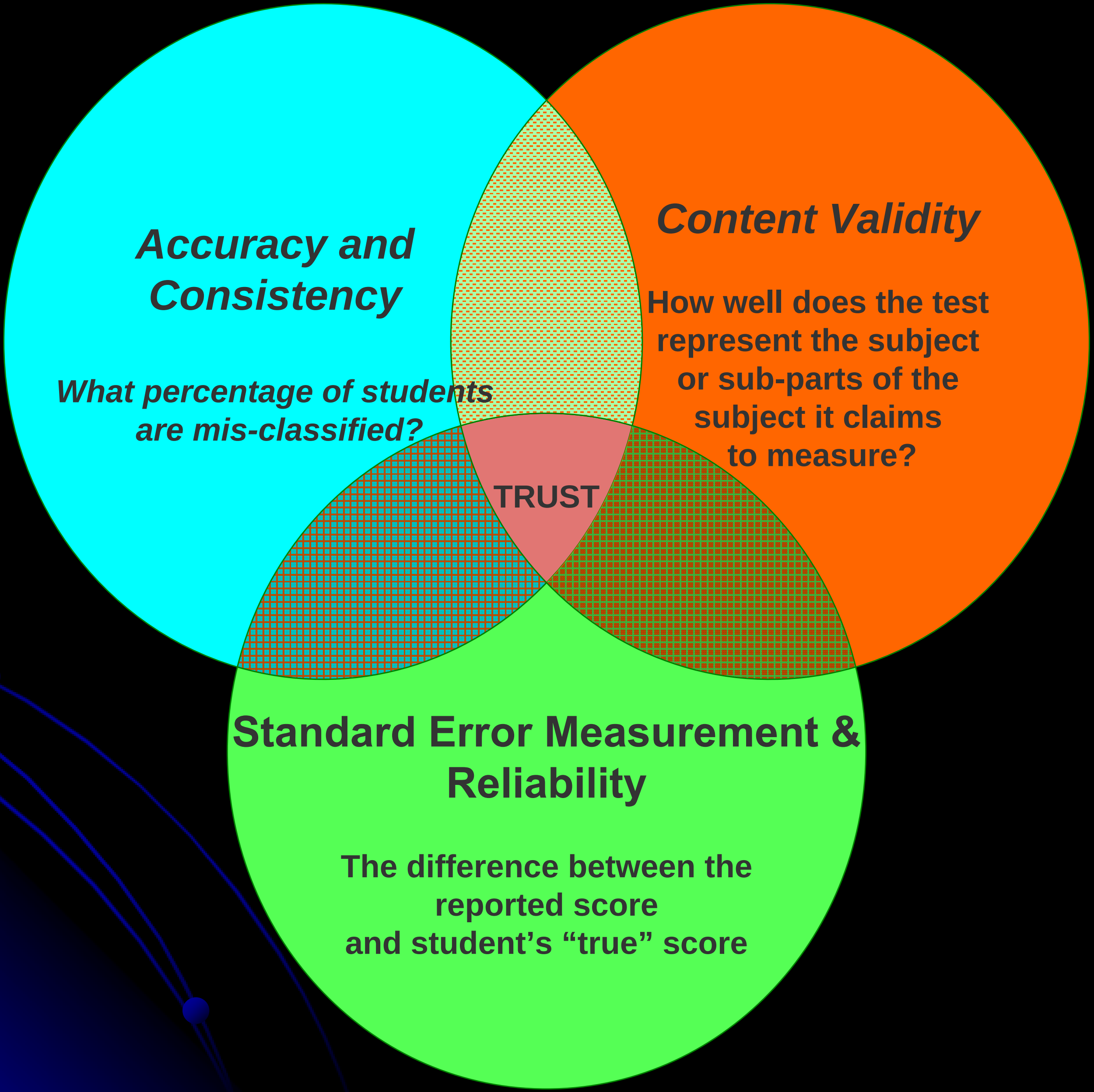
Limitations: So What?

1. The error present in state test results for individual students should cause us to be cautious about the way we use results to judge students. The score you see may not be the “true” score.

2. The levels of score reliability present in state tests should cause us to be cautious about making high stakes decisions about the effectiveness of curriculum and instruction.

3. The content coverage of the state tests should cause us to be cautious about the way we use results to make high stakes decisions about how much students know and can do and about the effectiveness of programs

Trust in Tests



Accuracy and Consistency

What percentage of students are mis-classified?

Content Validity

How well does the test represent the subject or sub-parts of the subject it claims to measure?

Standard Error Measurement & Reliability

The difference between the reported score and student's "true" score

TRUST

Trust in Tests: CAUTION

The technical qualities of the state test results do not support their use as the only factor when making high stakes decisions about students and programs.

Therefore:

- **District leaders should rethink the use of the state test results as the primary indicator of student achievement or program effectiveness.**

- **Leaders should create district-wide assessment practices to provide multiple tiers of achievement data that cover the spectrum of the NJCCCS and local program not assessed by state tests.**

OVERCOMING THE LIMITATIONS

Given the limits of state test results, consider developing:

- Programmatic eligibility based on multiple indicators

- Systematic District-wide Assessment Practices to augment the results from state tests

Programmatic Eligibility Criteria

- Example: Elementary Basic Skills
- The majority of the districts surveyed used the NJASK score of 199 or less as an automatic BSI placement indicator
- Given what we know about the error, score reliability, and content coverage…

Programmatic Eligibility Criteria

- Revise BSI criteria if you must use state test scores:

- Scores on NJASK of 190-210 – requires a review of the following as a minimum:

- Student achievement (Grades) as measured by high-quality classroom tests
- Teacher recommendation based on student achievement of classroom curriculum (district curriculum)
- Parent and/or student nomination or discussion
- Results from district-wide portfolio assessment system

Districts should make a judgment based on multiple factors to avoid misplacement-

OVERCOMING THE LIMITATIONS

District Assessment System

Systematic district-wide assessment practices to augment the results from state tests

What is an Assessment System?

- Multiple methods of measuring student growth and achievement against important learning targets

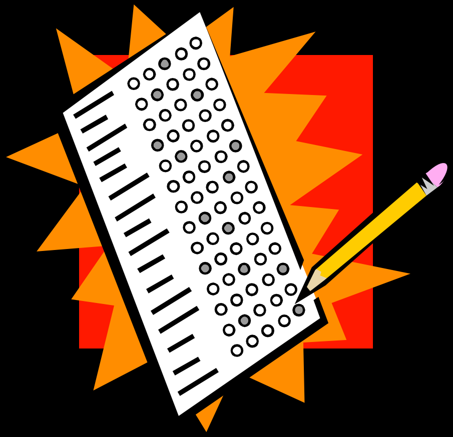
- Methods are related and part of a systemic initiative aimed at improving student learning and enhancing teaching

What is an Assessment System?

- Based on teachers developing high-quality, aligned assessments

- Calibration and Alignment Process

- Using a single interpretation of subject areas (common assessment strand)



Why an Assessment System?

- Acts as a running record or “portfolio” of achievement in relation to the district’s expanded curriculum

- Provide meaningful information relative to <u>within</u> year and year to year growth relative to high district standards

Multiple Measures

Assessment Examples:

- **Observation checklists**
- **Student portfolios**
- **Daily student work**
- **Classroom performance assessment**
- **Criterion referenced tests (State and other tests)**
- **Surveys**
- **Student interviews & conferences**

Assessment Framework (See Stiggins, 1994)

Method >>> To Assess: ∀	Selected Response	Essay / Graphic Org.	Interview/ Obs	Performance Assessment
Content Knowledge	Good Match	Good Match	Not the Best	Good Match
Problem Solving	Good for Some	Good for Some	Good for Some	Good for Some
Performance Skills	Not Good	Not Good	Good Match	Good Match
Use Skills to Create Products	Not Good	Not Good	Good Match	Good Match

K-8 Example

Grade Level	Skills Conferences (LA, Math, Social Skills)	Reading Attitude Survey	Writing portfolio Sep, Jan, May	Running Records and Reading Levels Pre-Post	NJASK / GEPA	NJPASS Math-LA	Science Open-Ended Pre / Post	School Attitude Surveys	Focus Group Interview	Pre & Post Math Open-Ended
K	X		X	X						
1	X		X	X						
2	X		X	X		X				
3	X		X	X	X		X			X
4			X		X		X			X
5		X	X		X		X			X
6		X	X		X		X	X		X
7		X	X		X		X	X	X	X
8		X	X		X		X	X	X	X

Assessment Methods

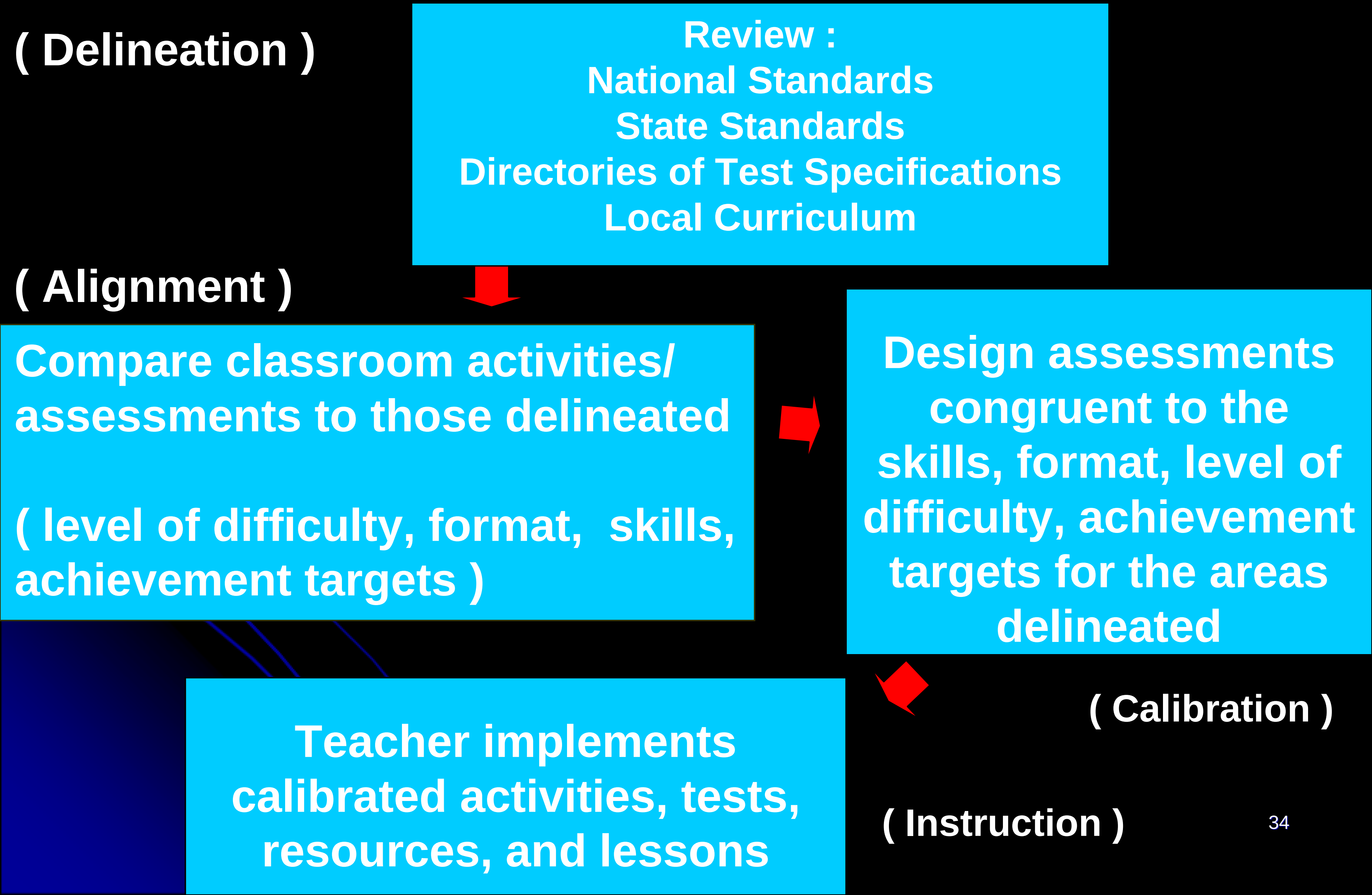
Calibration is Crucial !!!

- WE MUST look at the activities and instructional strategies used -
  - *Are they aligned with the*

    skills and knowledge - level of difficulty - cognitive format …

    required by the NJCCCS, NJASK, GEPA and your LOCAL curriculum?

Classroom Calibration and Alignment



(Delineation)

Review :
National Standards
State Standards
Directories of Test Specifications
Local Curriculum

(Alignment)

Compare classroom activities/ assessments to those delineated

(level of difficulty, format, skills, achievement targets)

Design assessments congruent to the skills, format, level of difficulty, achievement targets for the areas delineated

(Calibration)

Teacher implements calibrated activities, tests, resources, and lessons

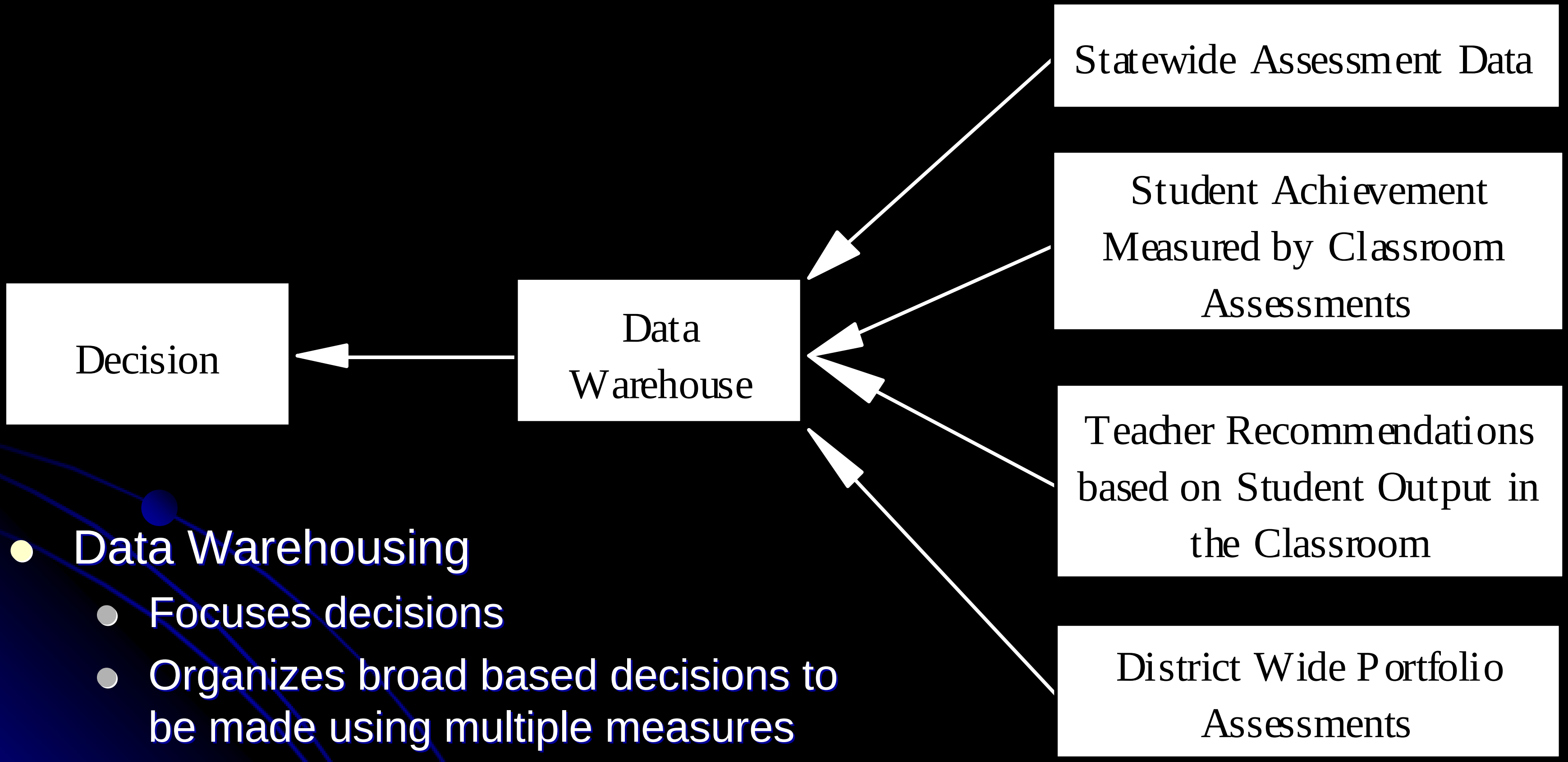
(Instruction)

OVERCOMING THE LIMITATIONS

- Requires developing multiple measures to view student achievement through multiple lenses.

- Multiple, high-quality, measures will provide the type of data needed to make high quality decisions about students.

Multiple Measures via Data Warehousing



Statewide Assessment Data
Student Achievement Measured by Classroom Assessments
Teacher Recommendations based on Student Output in the Classroom
District Wide Portfolio Assessments
Data Warehouse
Decision

- Data Warehousing
  - Focuses decisions
  - Organizes broad based decisions to be made using multiple measures
  - Provides longitudinal decision making opportunities over time

OVERCOMING THE LIMITATIONS

Our students have a right to a quality education…

We have a duty to provide it.

Thank You

- Feel free to contact us for more information or assistance

- Christopher Tienken, EdD

  goteach1@hotmail.com

- Patrick Michel, EdD

  pmichel@hhsd.k12.nj.us

- Thomas Tramaglini, EdM

  ttram@freeholdboro.k12.nj.us