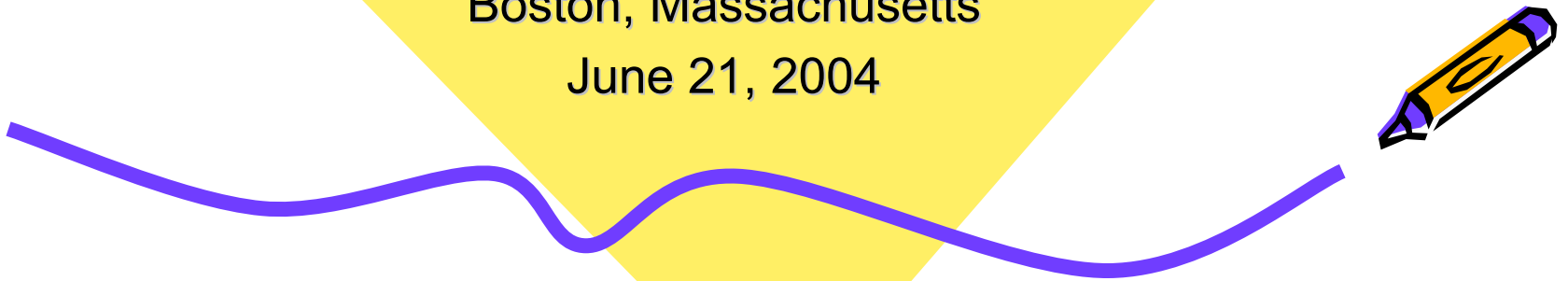
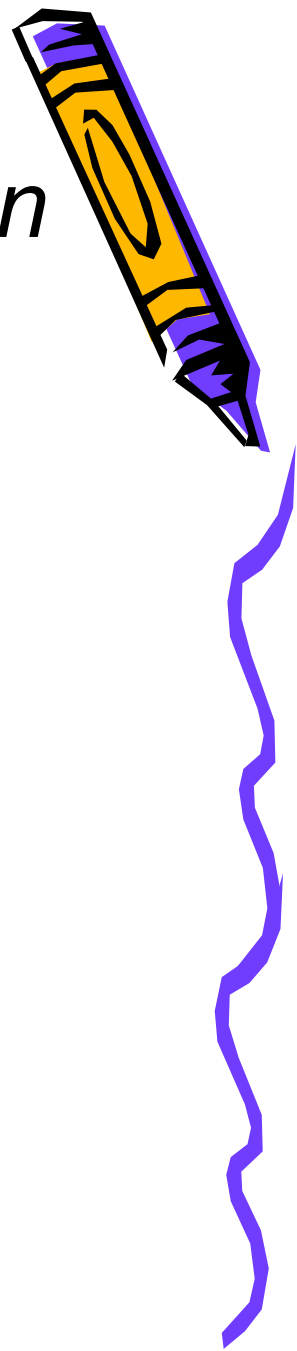


*Setting Standards
Across Grades 3 through 8:
It is More than Choosing the Right Method*

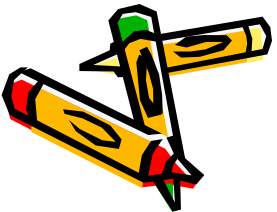
CCSSO Large-Scale Assessment Conference
Boston, Massachusetts
June 21, 2004



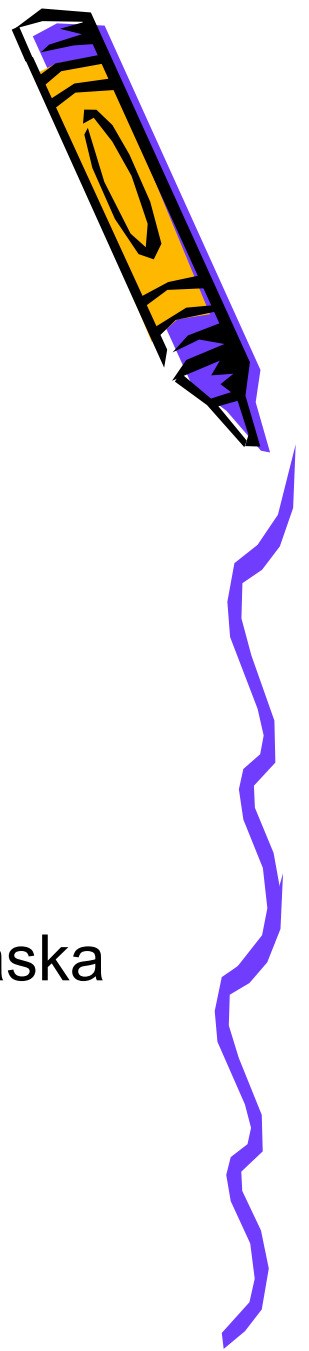
*Standard Setting is a journey down
a long and winding road...*



*...and assigning cut points on a
test is a final step in the process*

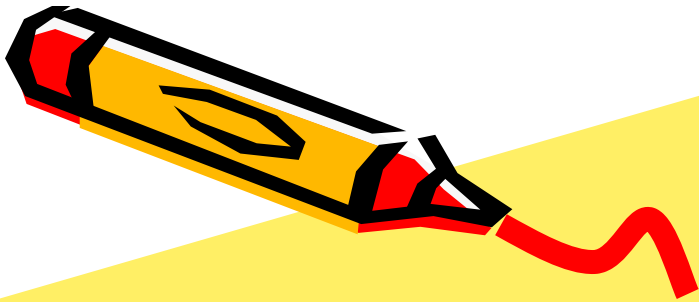


Presenters



- **Jeff Nellhaus**
 - Massachusetts Department of Education
- **Charles DePascale**
 - National Center for the Improvement of Educational Assessment
- **Barbara Plake**
 - Buros Center for Testing, University of Nebraska Lincoln
- **Mike Beck**
 - BETA, Inc.





Standard Setting is More than Choosing a Method

Jeff Nellhaus
Massachusetts Department of Education

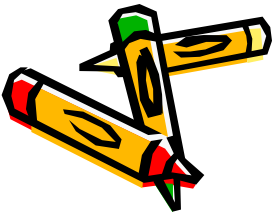
National Large-Scale Assessment Conference
June 21, 2004

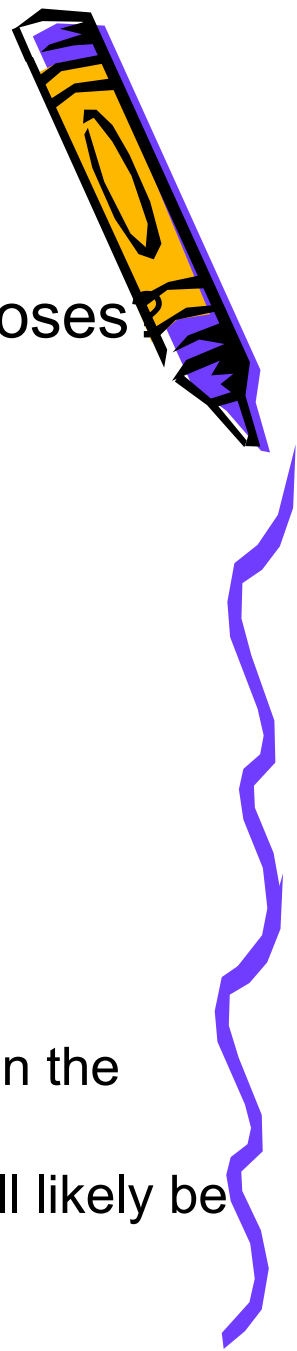


Setting Performance Standards



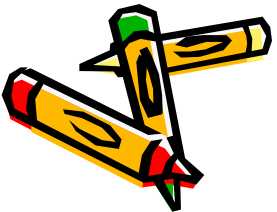
1. Determining the purpose(s) the standards will serve
2. Determining the number of performance levels and the names of the levels
3. Describing the levels
4. Developing a test aligned with the standards
5. Determining performance level cut scores



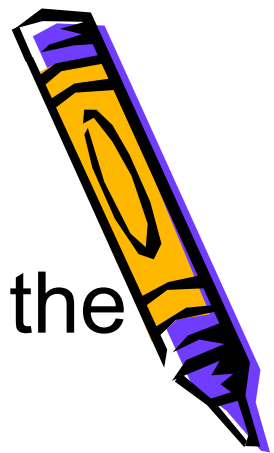


1. Determining the purpose(s) the standards will serve

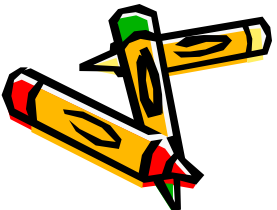
- Will the standards be used for one or multiple purposes?
 - Student determinations
 - Grade promotion
 - Graduation
 - Remediation
 - School or program determinations
 - Improvement - Adequate Yearly Progress
 - Underperforming schools
 - Grant allocations
- Issues:
 - Without a clear purpose in mind, the remaining steps in the process will suffer.
 - Even if the purpose is well identified, the standards will likely be used for unintended/unanticipated applications.



2. Determining the number of the levels

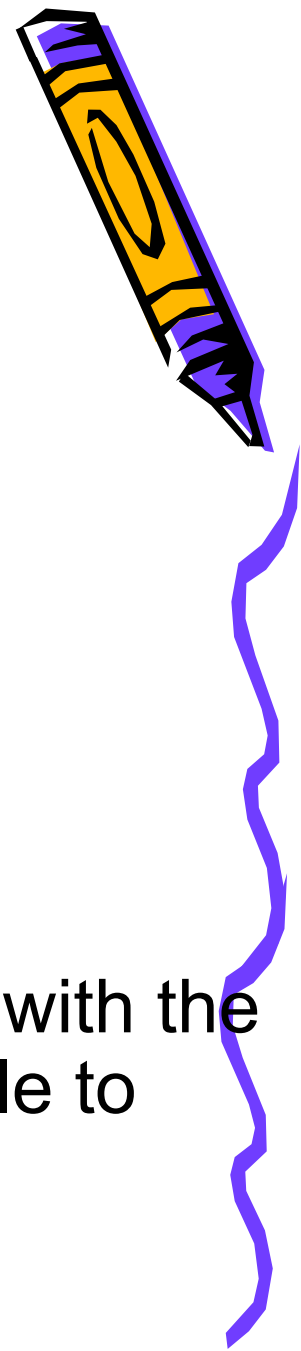


- The number of the levels should follow the purpose(s)
 - Single purpose tests for making student determinations may require only two levels
 - Tests used to measure whether schools are improving may require multiple levels
- Issue: The number of levels should not exceed the capacity of the test to differentiate among them



2. Determining the names of the levels

- Naming the levels tends to be controversial
 - Stay neutral?
 - Level I, Level II, Level III, Level IV
 - Mirror NAEP?
 - Advanced, Proficient, Basic, Below Basic
 - Send a message?
 - Advanced, Proficient, Needs Improvement, Warning
 - Split hairs?
 - Advanced, Proficient, Basic, Below II, Below I
- Issue: The names that should be consistent with the purpose/uses of the standards and acceptable to various audiences



3. Describing the performance levels



- **Generic descriptions**
 - Applicable to any content area
 - May describe general level of knowledge and skills at each level (e.g., partial, strong, in-depth)
 - May describe the consequence of attaining the level
- **Content-specific descriptions**
 - Applicable to a particular content area
 - Describe the specific skills and knowledge required to attain the level
 - May be broken down further by grade or grade span
- **Issues**
 - Incorporating content standards
 - Avoiding in the descriptions performances not measured by the assessment system
 - Writing descriptions that are coherent across grade levels



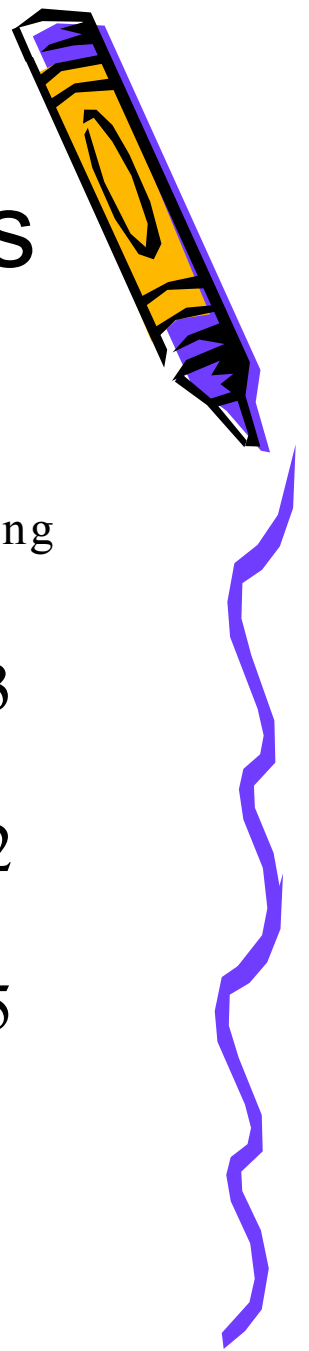
4. Ensuring that the test is aligned with the standards

- Optimally, tests are selected or developed *after* performance level descriptions are written
- There must be a sufficient number of items to make consistent and accurate determinations at each of the levels
- There must be items on the test that provide students the opportunity to demonstrate skills and knowledge called for in the performance level descriptions
- Issue (example):
 - In MA, the grade 4 ELA test initially was not well-aligned with the grade 4 ELA performance standards. The standards required that students at the Proficient level be able make comparisons between two stories. However, there were no tasks on the assessment requiring them to do so. As a result ...

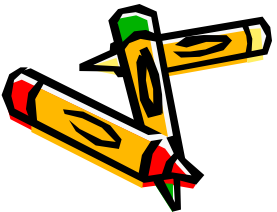


MCAS Grade 4 ELA Results

(percentage of students)



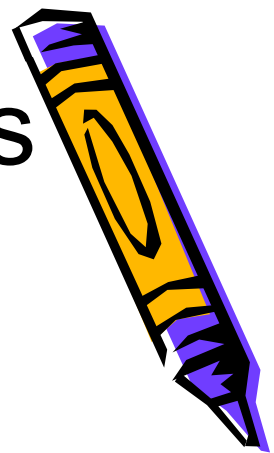
	Advanced	Proficient	Needs Improvement	Failing
2000	1	19	67	13
1999	0	21	67	12
1998	1	19	65	15



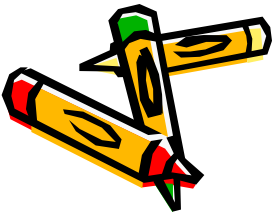


MCAS Grade 4 ELA Results

(percentage of students)

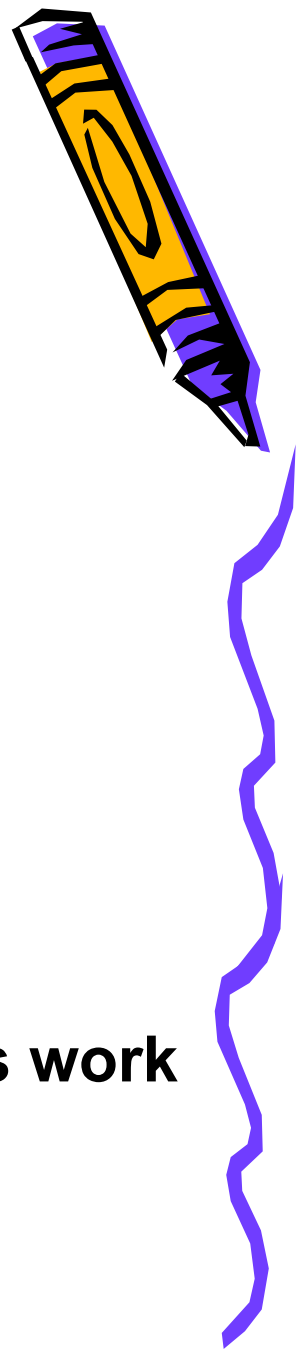


	Advanced	Proficient	Needs Improvement	Failing
2001	7	44	38	11
2000	1	19	67	13
1999	0	21	67	12
1998	1	19	65	15

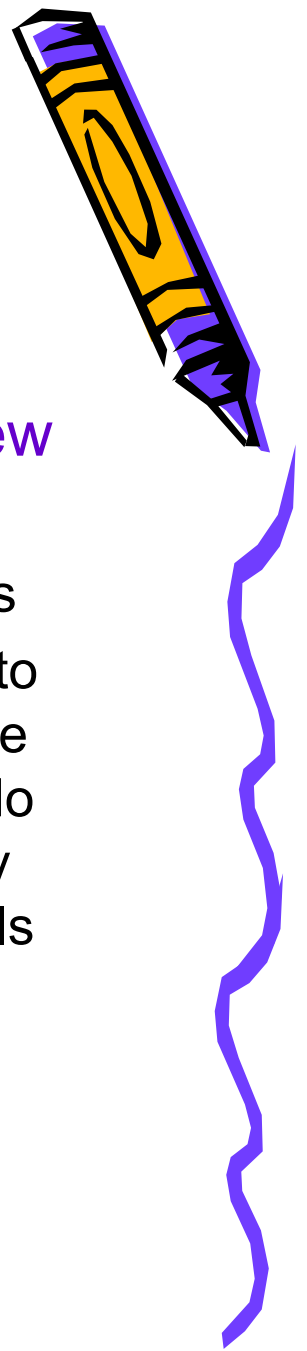


5. Determining performance level cut scores

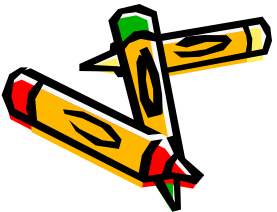
- **Selecting the method**
 - Test centered or student centered?
- **Selecting panel members**
 - How many?
 - Educators only? If non-educators, why?
 - Diversity? What kind? Who chooses?
- **Determining the number of rounds of judgments, feedback provided to panelists after each round**
 - None?
 - Rater feedback only? Impact data? If so, what kind?
- **Issue: Finalizing cut scores derived from panel's work**
 - On what basis can adjustments be justified?



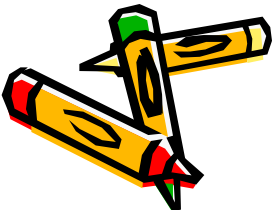
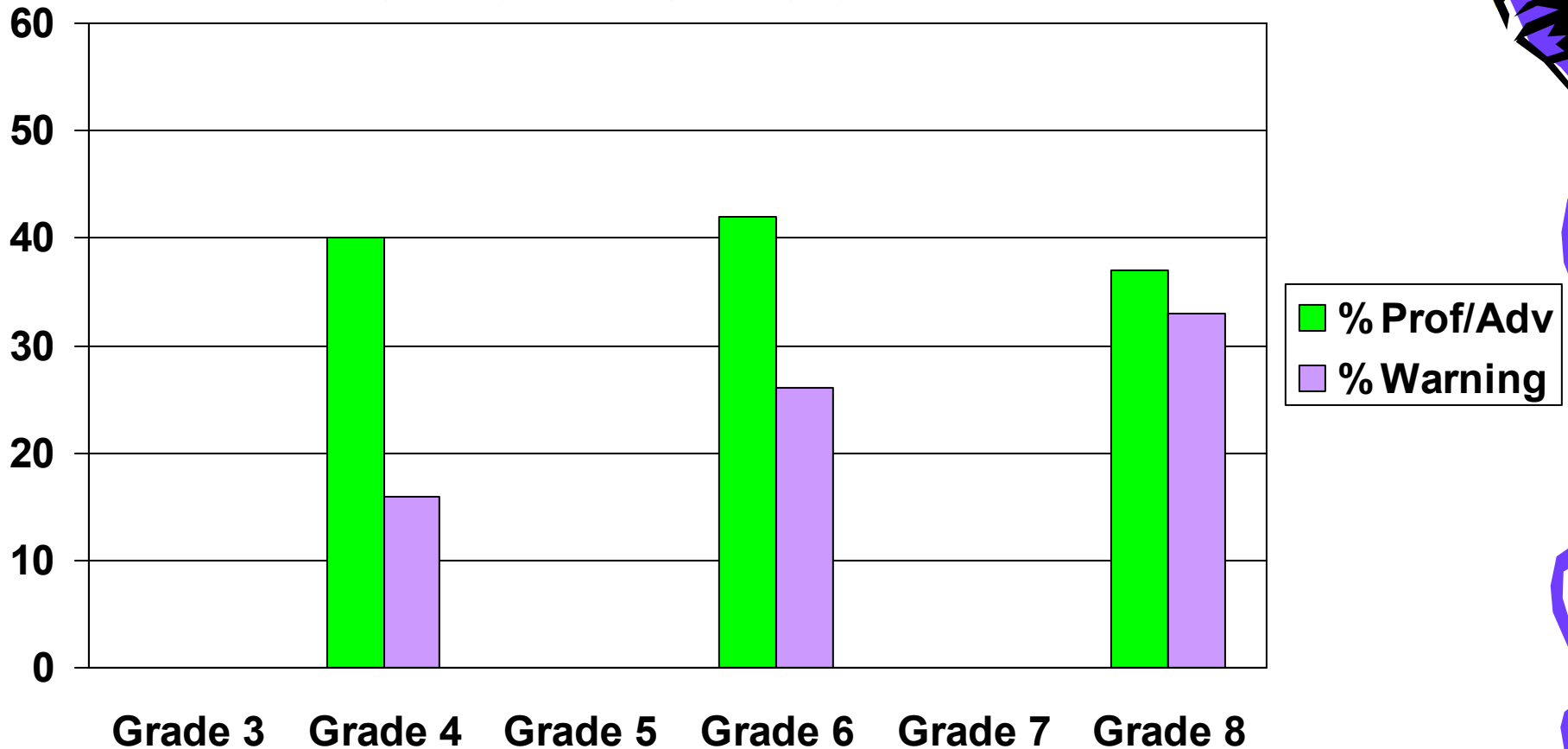
Issues States Will Face as they Add Tests to Comply with NCLB



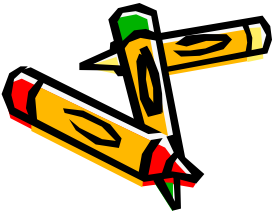
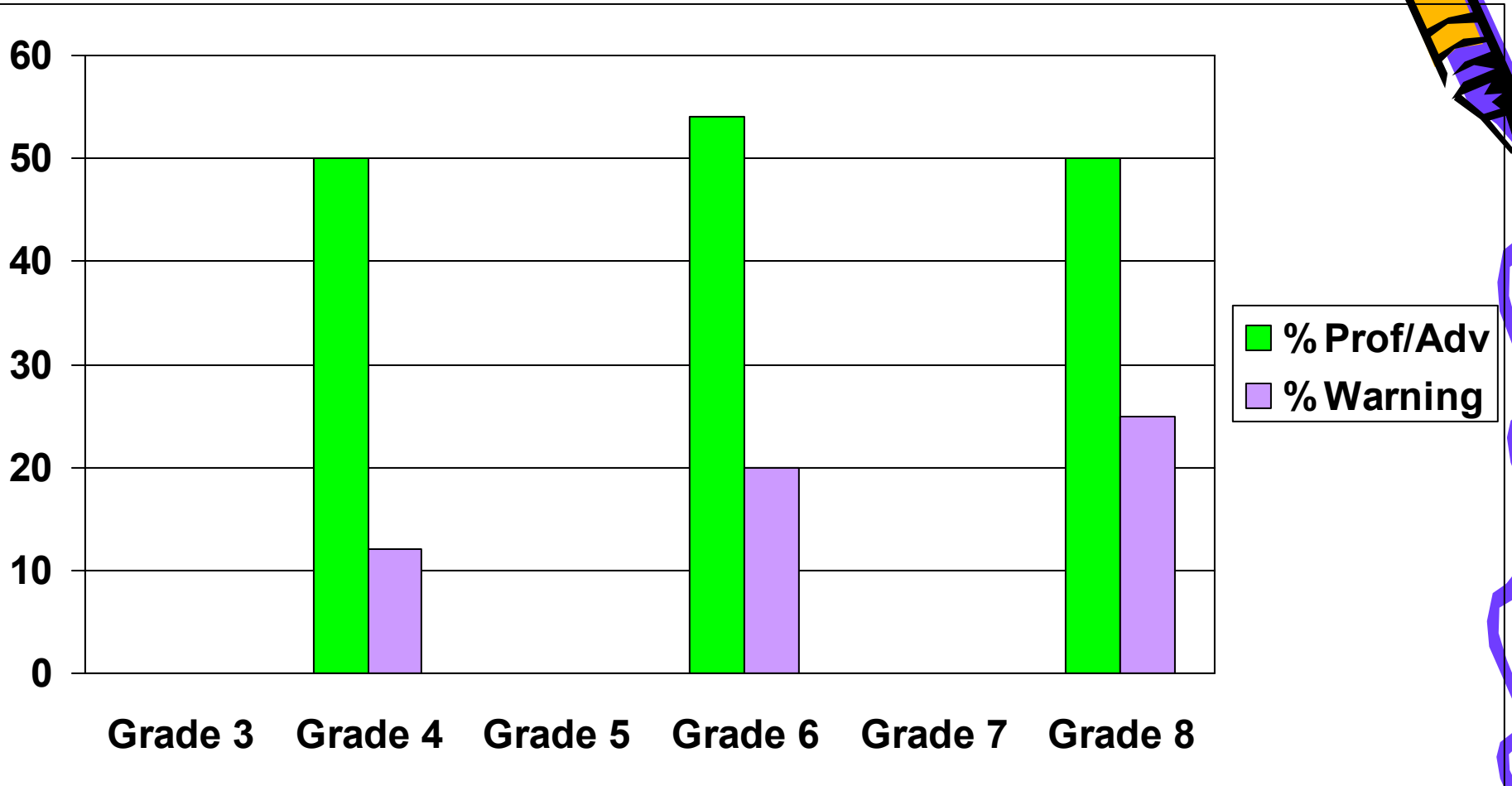
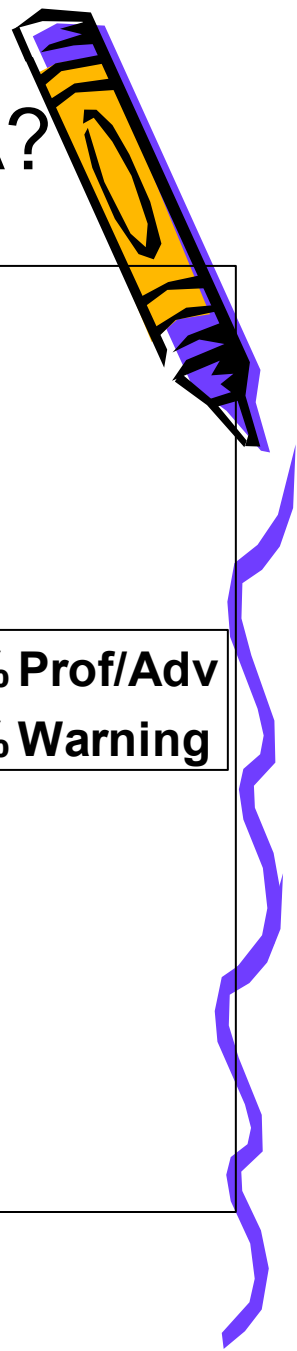
- With respect to existing tests
 - Maintain standards
 - Maintain trends
 - Only if content, performance standard, test specifications remain the same
 - Set new standards
 - Must do, if content or performance standards changed, test specs changed
 - May want to do to correct existing problems (e.g., current results seem too high or low, inconsistent across grades or grade spans)
- With respect to new tests
 - Set new standards
 - Challenge will be to report performance level results that do not fluctuate wildly across grade levels for unexplainable reasons



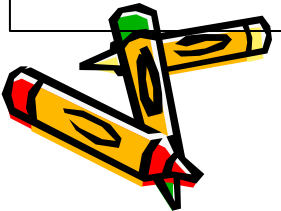
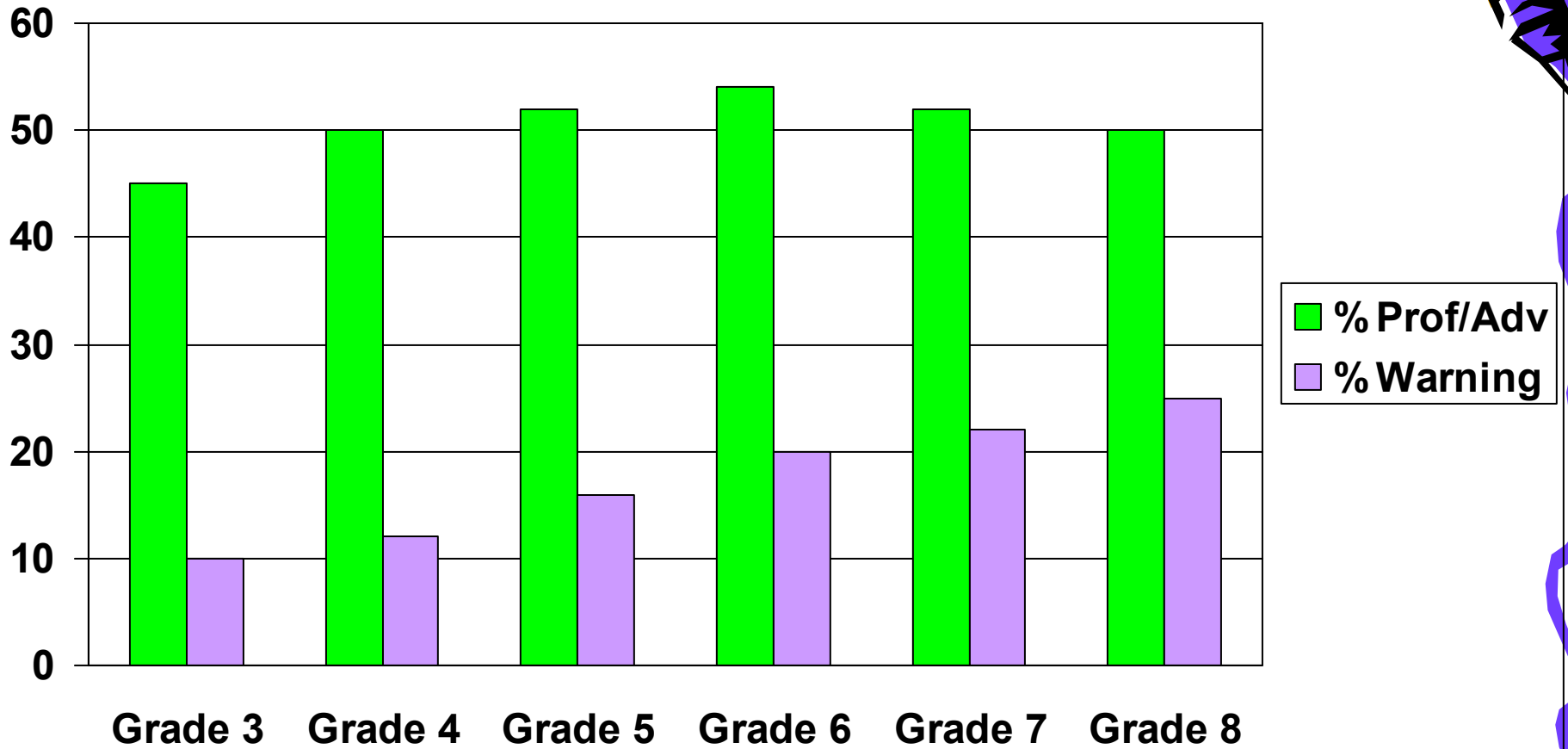
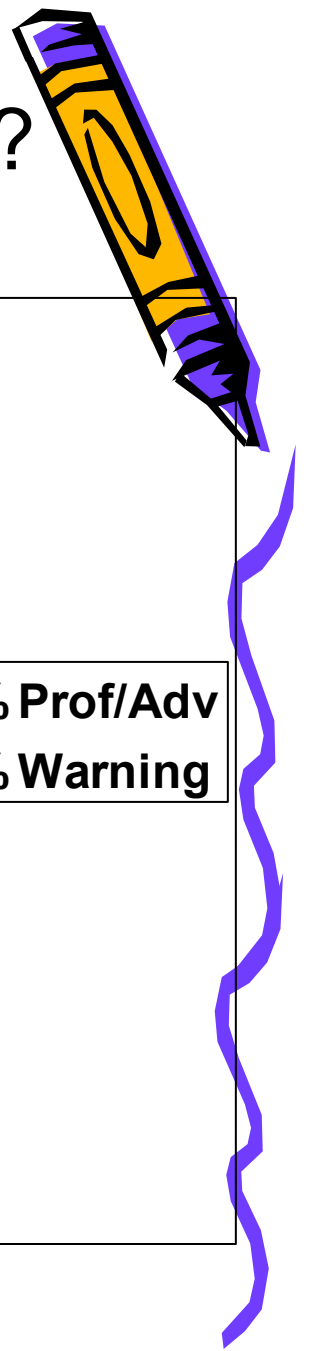
2003 Mathematics Performance in MA



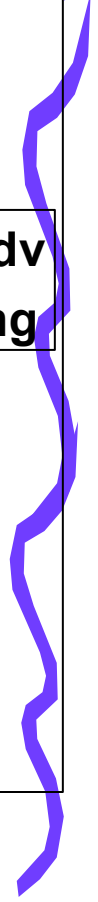
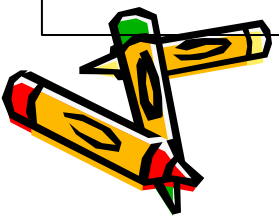
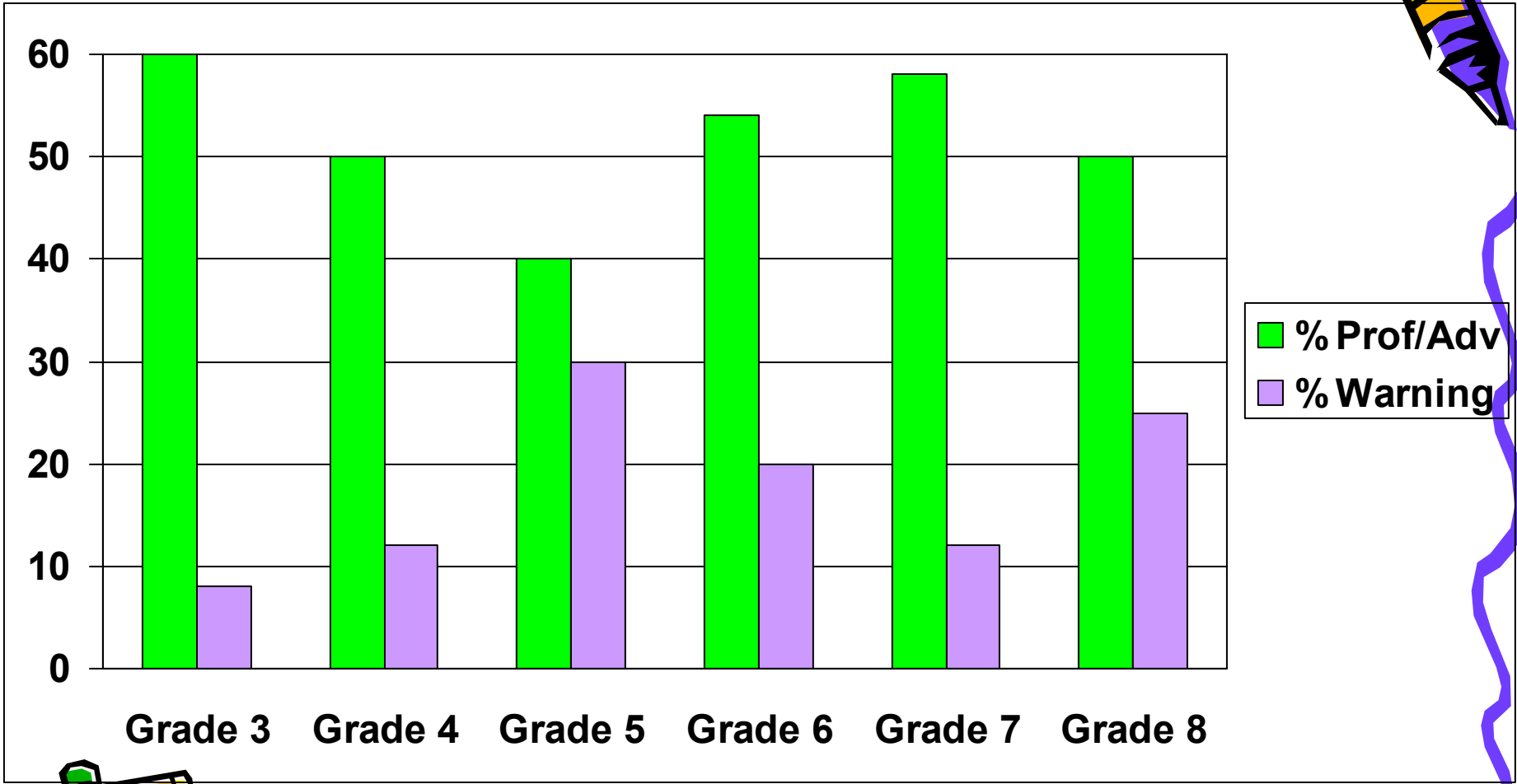
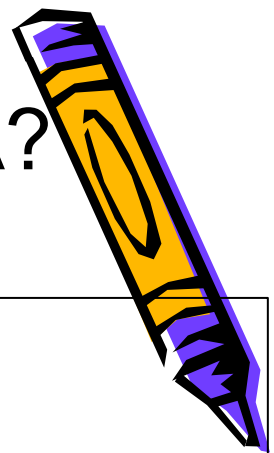
2006 Mathematics Performance in MA?

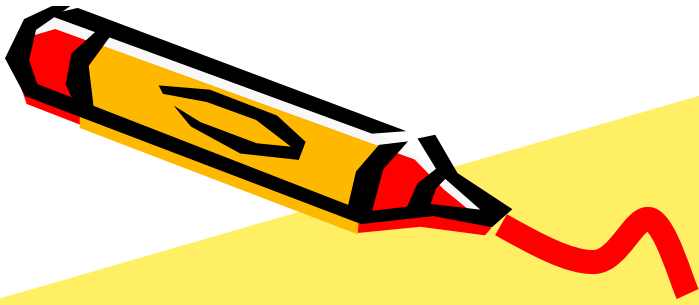


2006 Mathematics Performance in MA???



2006 Mathematics Performance in MA?





Developing Content Standards and Defining Performance Levels

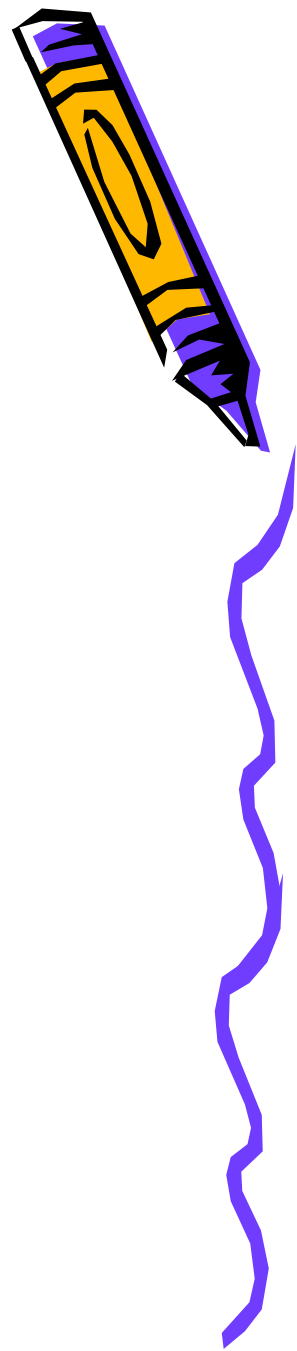
Charles A DePascale

National Center for the Improvement
of Educational Assessment



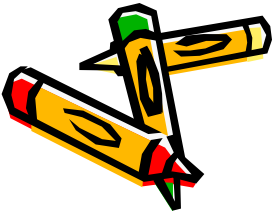


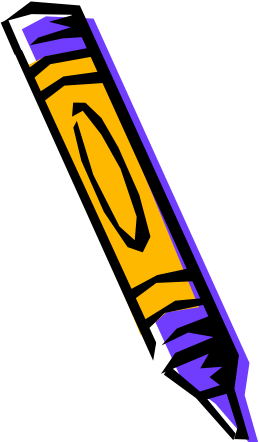
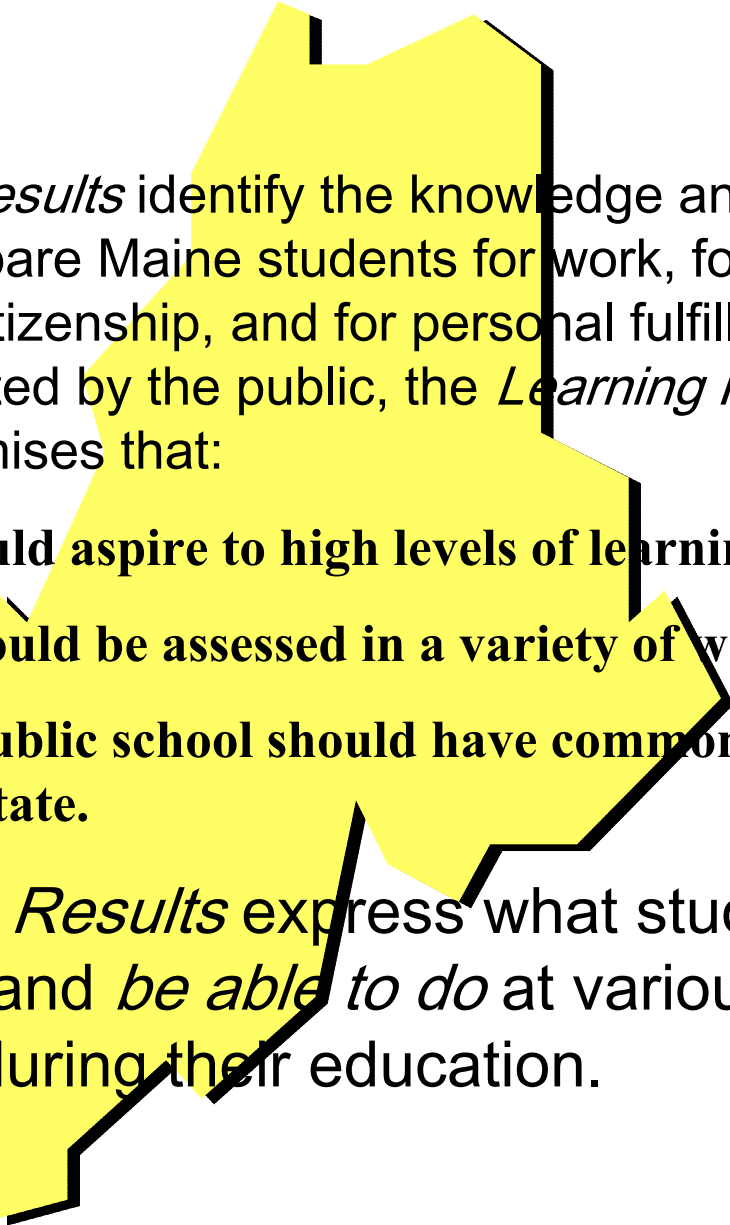
In the beginning



- ??? created content standards.
- The Board approved them.
- And it was good.

But what did the standards represent?







The *Learning Results* identify the knowledge and skills essential to prepare Maine students for work, for higher education, for citizenship, and for personal fulfillment. Strongly supported by the public, the *Learning Results* are built on the premises that:

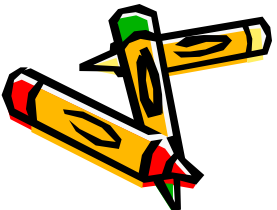
- **all students should aspire to high levels of learning;**
- **achievement should be assessed in a variety of ways; and**
- **completion of public school should have common meaning throughout the state.**


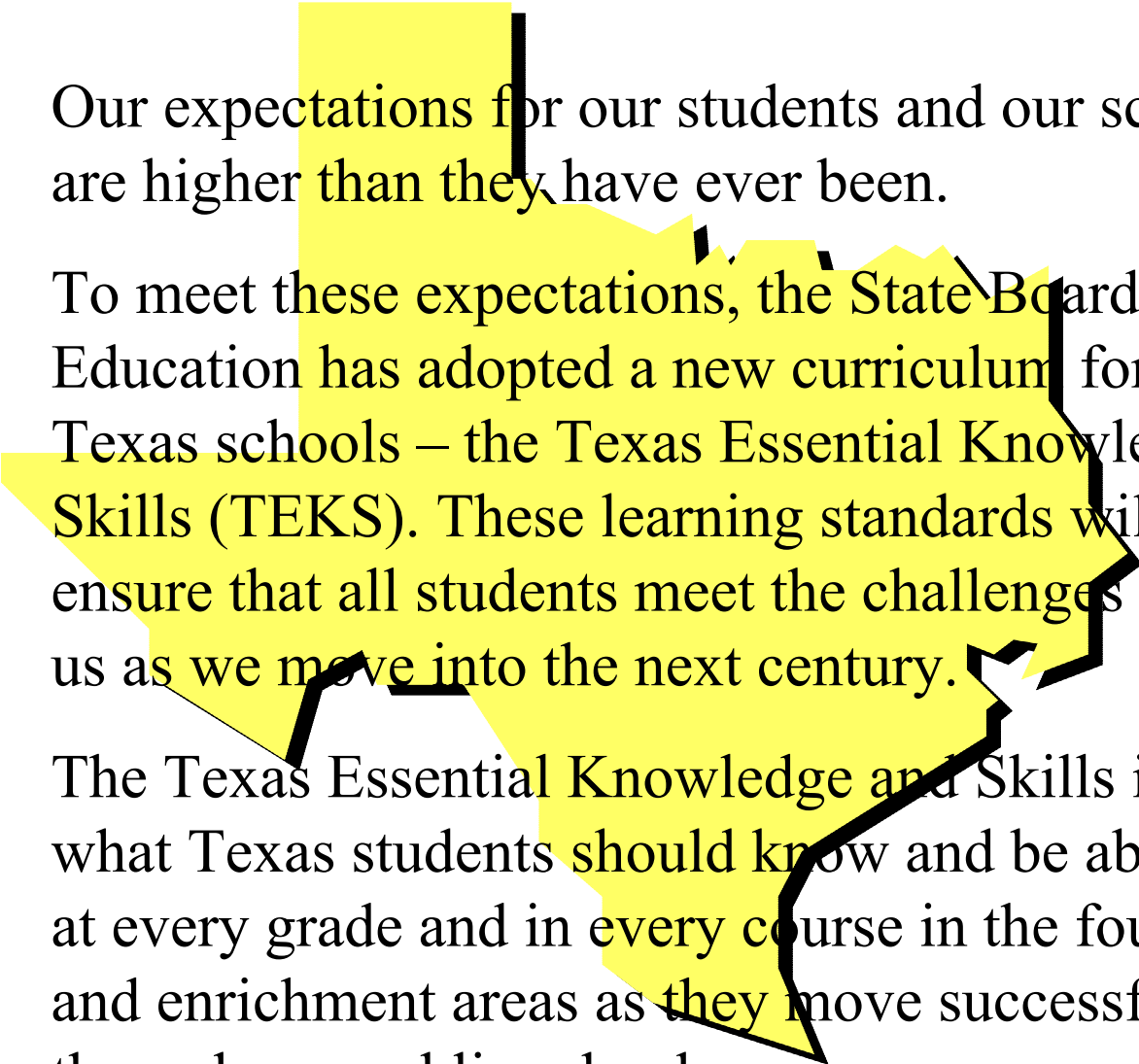
The *Learning Results* express what students *should know* and *be able to do* at various checkpoints during their education.



Content standards were developed to encourage the highest achievement of every student by defining the knowledge, concepts, and skills that students should acquire at each grade level.

With the adoption of these content standards in English-language arts, California is setting a new form. We are redefining the state's role in education. For the first time, we are stating - explicitly - the knowledge and skills that students need to acquire at each grade level through the end of grades nine and ten and grades eleven and twelve.

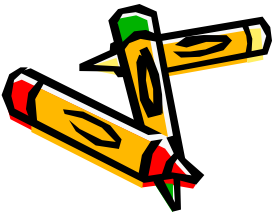




Our expectations for our students and our schools are higher than they have ever been.

To meet these expectations, the State Board of Education has adopted a new curriculum for all Texas schools – the Texas Essential Knowledge and Skills (TEKS). These learning standards will help us ensure that all students meet the challenges ahead of us as we move into the next century.

The Texas Essential Knowledge and Skills identify what Texas students should know and be able to do at every grade and in every course in the foundation and enrichment areas as they move successfully through our public schools.

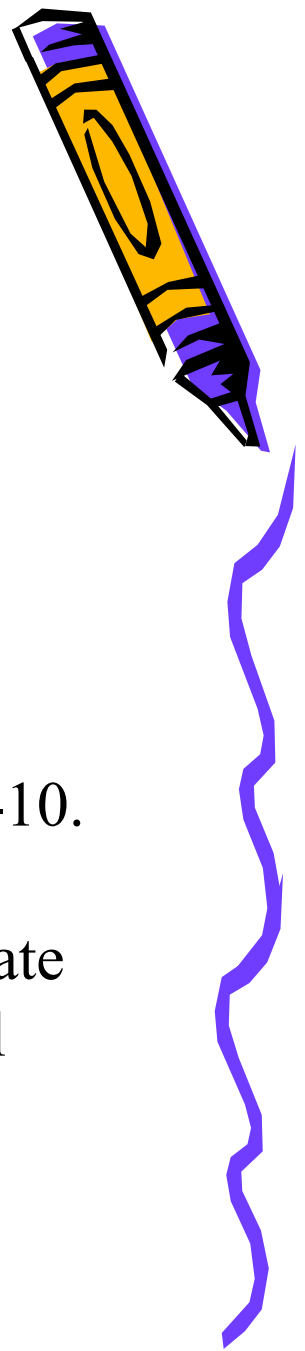


Content Standards

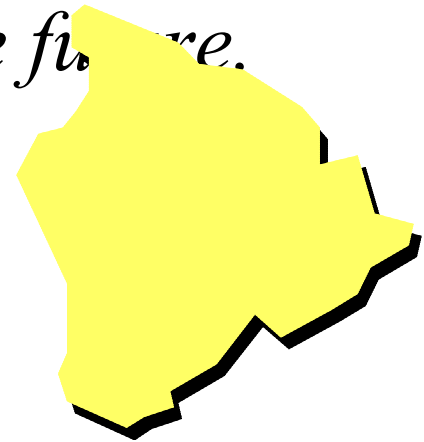
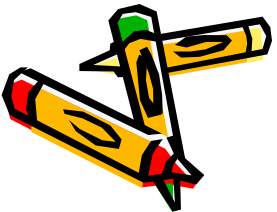
The document's *content standards* define what every Delaware student should know and be able to do.

Performance Indicators

Each *content standard* is followed by *performance indicators* in grade-level clusters: K-3, 4-5, 6-8 and 9-10. Students in Delaware will be assessed using these indicators. The *performance indicators* clearly articulate the specific expectations of students in the grade-level clusters from kindergarten through tenth grade.



....Our standards set the course,
while students, families, and
community fill the sails with
expectation as we voyage with the
treasure of bright, young minds
ready to lead the way to the future.

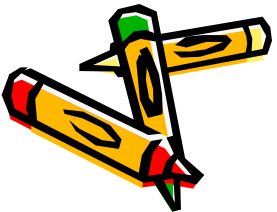


Content Standards define *Proficient*



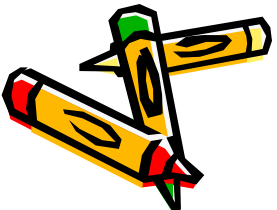
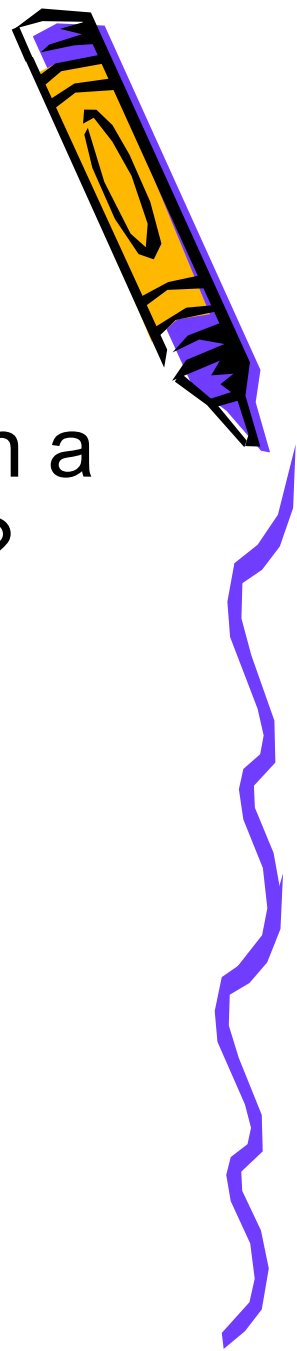
- In many states, the content standards define the performance that is expected of all students – proficient.

An important part of standard setting is complete when the content standards are adopted.



But there's more...

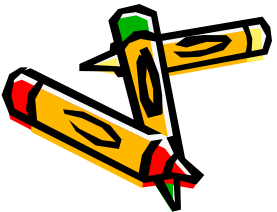
- What does proficient performance on a particular content standard look like?
 - Content standards begat performance indicators.
 - Performance indicators begat ...

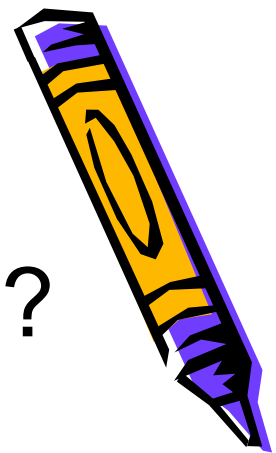


A. NUMBERS AND NUMBER SENSE

Students will understand and demonstrate a sense of what numbers mean and how they are used.

- 1. Apply concepts of ratios, proportions, percents, and number theory (e.g., primes, factors, and multiples) in practical and other mathematical situations.*
- 2. Compute and model all four operations with whole numbers, fractions, decimals, sets of numbers, and percents, applying the proper order of operations.*

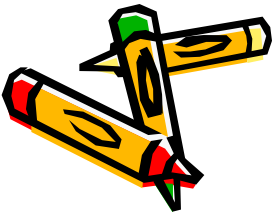




So, how good is good enough?

- Does a student have to demonstrate mastery of all of the content standards and performance indicators?
- Does a student have to demonstrate mastery of most of the ...?
- Does a student have to demonstrate mastery of any of the...?

Answering these questions is another major step in the standard setting process.

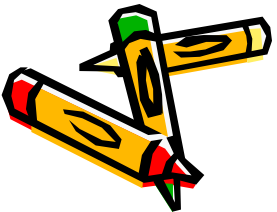




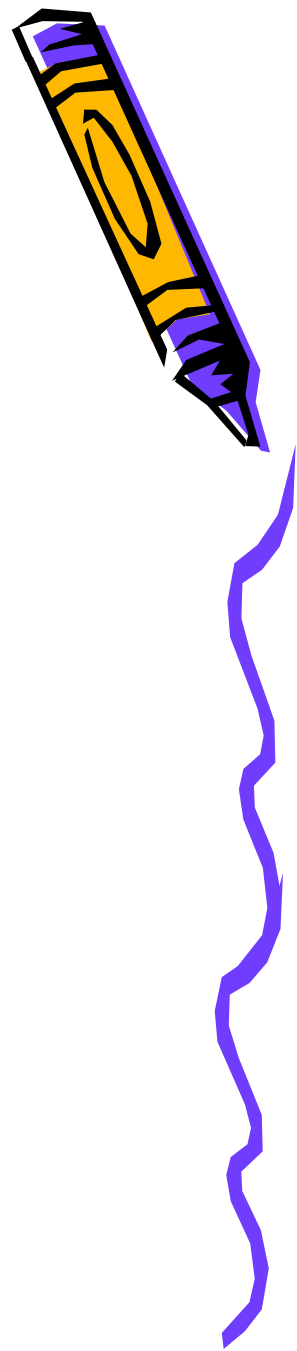
Let there be more levels.



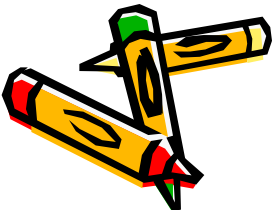
- Proficient is not sufficient.
- More performance levels are needed:
Advanced, Basic, Below Basic,
- Can this be good?



Defining the additional performance levels



- The content standards defined only proficient performance?
- How do you define Basic and Advanced performance?
 - *Basic is less than proficient*
 - *Advanced is more than proficient*



NAEP Achievement Levels

“How good is good enough”



Basic

This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.

Proficient

This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter.

Advanced

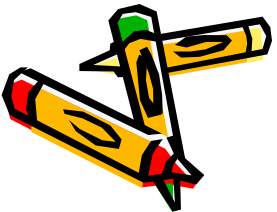
This level signifies superior performance.



Partial Mastery?



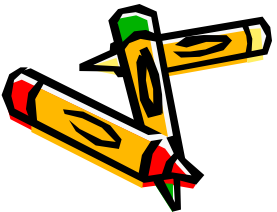
- Does partial mastery mean
 - mastery of some of the content standards and not others?
 - mastery of “part” of all of the content standards?
 - performance at a cognitive level that is less than mastery?



Moving down the standard setting road



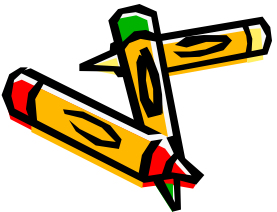
- Deciding how to define basic and advanced performance in relation to proficient performance.
- Writing descriptions of performance at the basic and advanced levels (content standards and performance indicators)

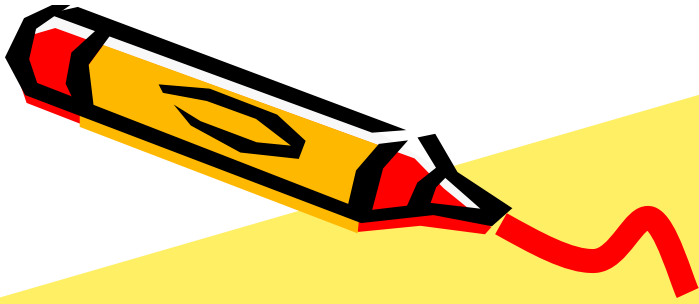




nd NCLB said,

Let the states create tests to measure student performance against the content standards and performance standards. Let the states use those tests results to classify student performance into a performance level. And it was so.





Role of Alignment in the Standard Setting Process

Barbara S. Plake

Buros Center for Testing

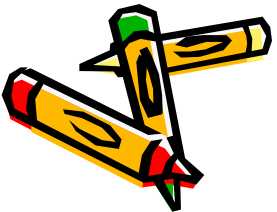
University of Nebraska-Lincoln



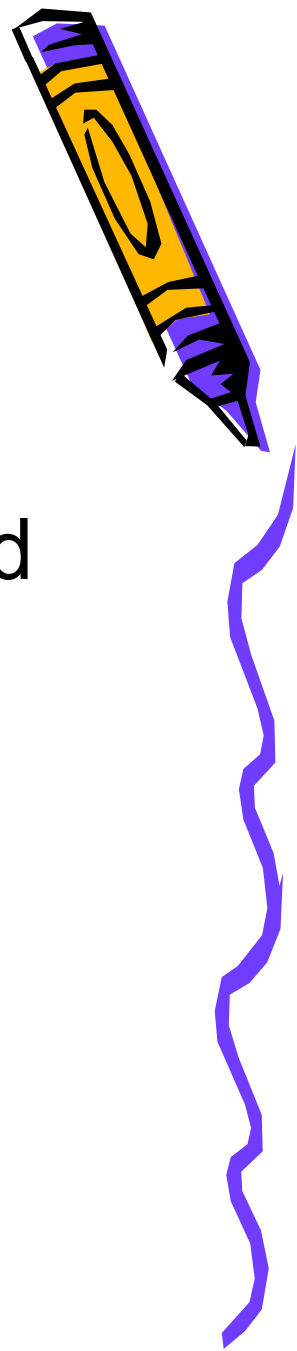
Why Is Alignment Part of the Standard Setting Process?



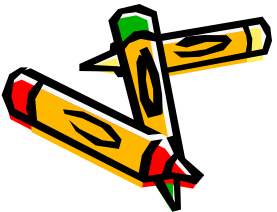
- Standard setting is used to enhance score interpretation
- Performance on the test is used to assign scores to performance categories
- In order for performance interpretations to be meaningful, scores must be linked to what the test is intended to measure



Linkage Between Score Interpretation and Standard Setting Results



- Test should be aligned to content specifications (content standards and content weightings)
- Test should be designed to provide sufficient data points to allow for meaningful performance category interpretations



Dimensions of Alignment

- More than just linkage to content specifications
- Must ensure content matches intended construct
 - Cognitive complexity
 - Performance level expectations

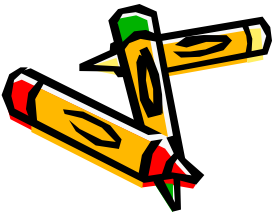
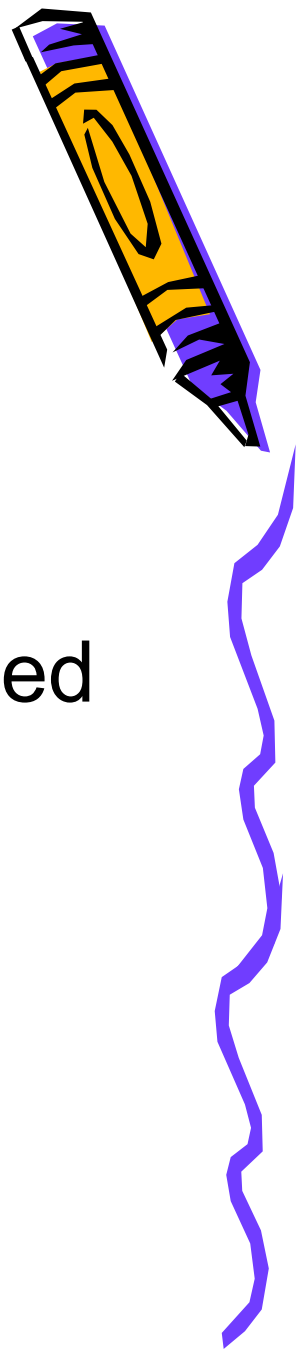


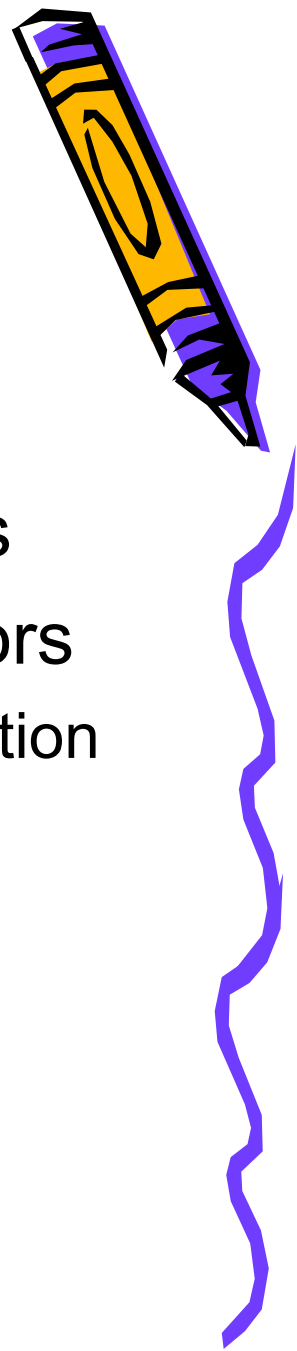
Illustration of Alignment Analysis



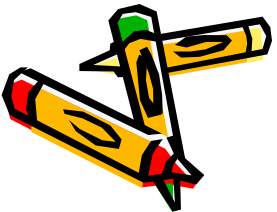
- Convene a panel of practitioners
- Training on
 - Content specifications, including content components (standards) and content weightings
 - Performance level descriptors



Tasks for Alignment Process

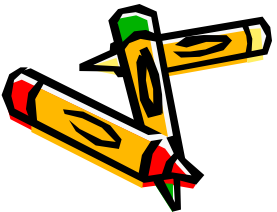
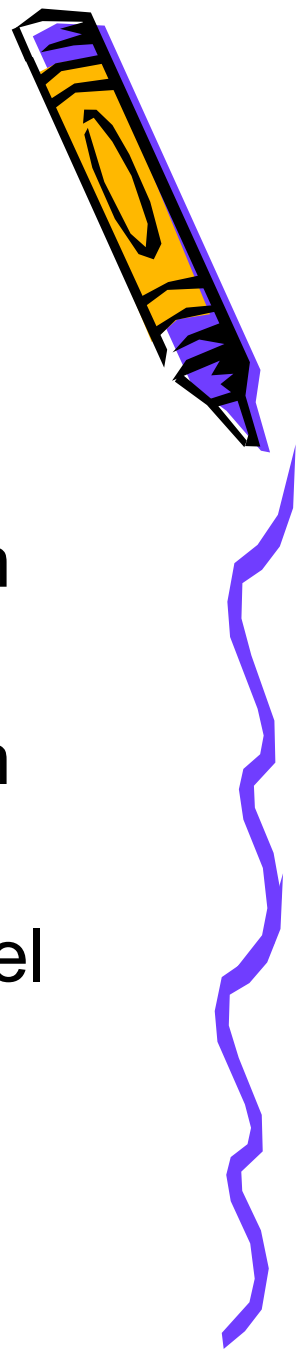


- Discussion of content specifications, including cognitive complexity dimensions
- Discussion of performance level descriptors
 - Differences in depth of understanding/application
 - Differences in substance
 - Differences in amount

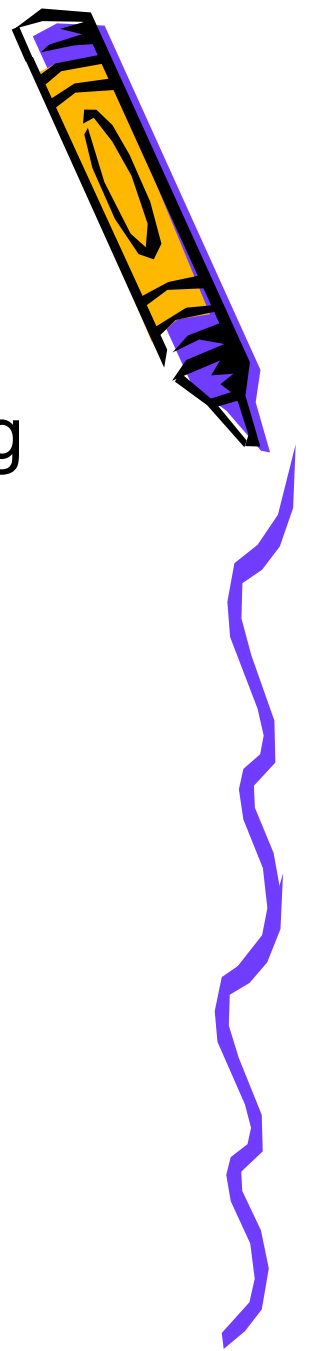


Rating Forms

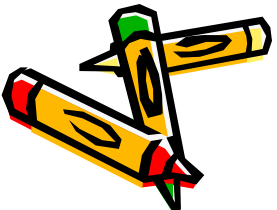
- Form provides for three ratings
 - Degree to which test tasks/items match content specification
 - Degree to which test tasks/items match cognitive complexity
 - Assignment of item to performance level descriptor



Example of Rating Form

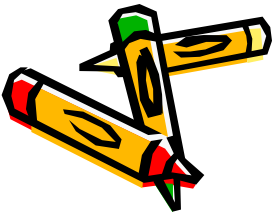
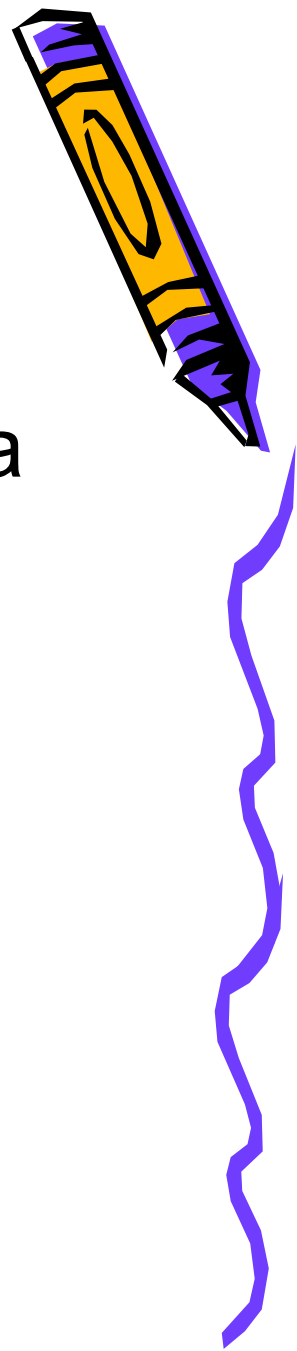


- Statewide assessment program in reading and mathematics
- Customized test
- Performance level descriptors
- Rating process
 - Rating of primary match
 - Rating of cognitive level
 - Performance category assignment



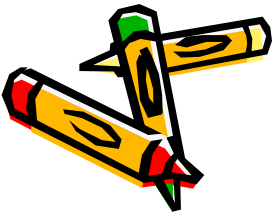
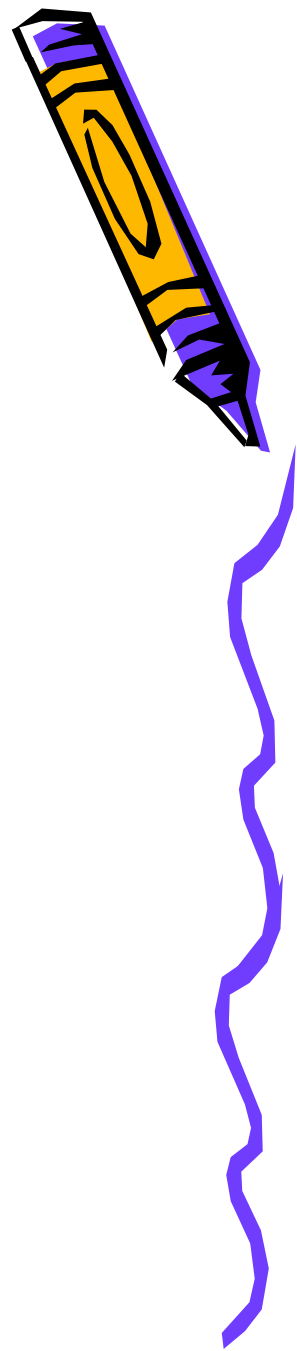
Procedure

- Grade level teachers by content area
- Multiple panels at each grade level
- Independent assignment
- Group discussion
- Consensus rating



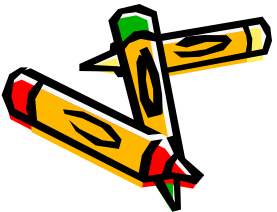
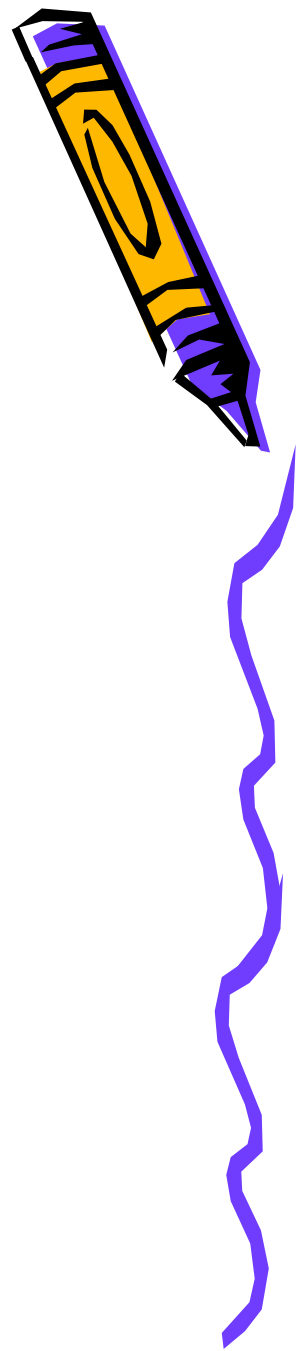
Analysis

- Average ratings by item for
 - Match
 - Cognitive Complexity
 - Performance Category



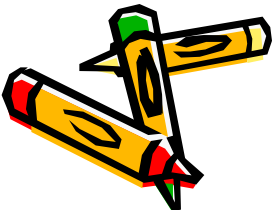
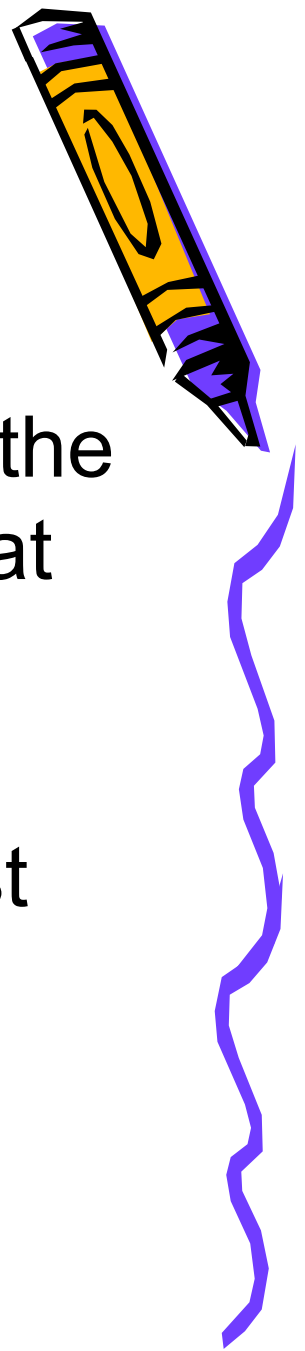
Aggregated Ratings

- By content area across grade levels
 - Average across panels
 - Match
 - Cognitive Complexity
 - Performance Category



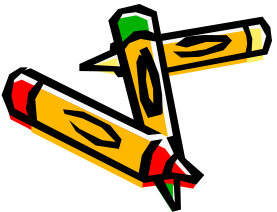
Interpretation

- Expect at least a 75% agreement in the match to content specifications and at least 5 - 7 items/points for each performance category
- May suggest needed revisions in test components

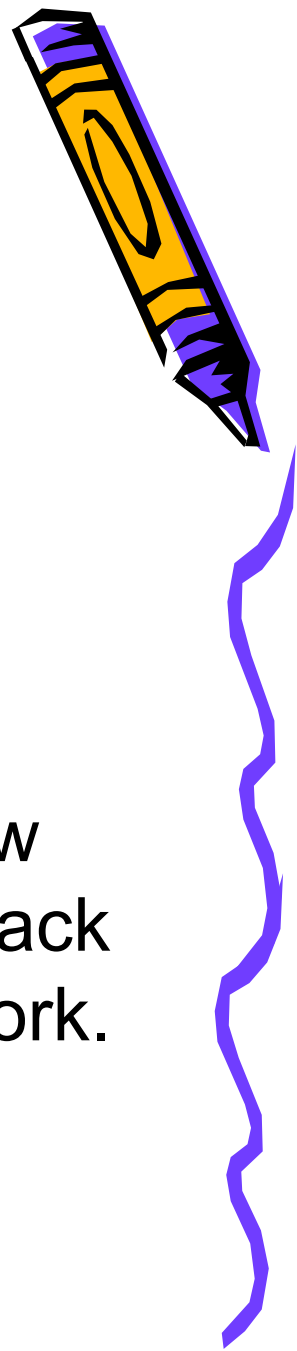


Variations in Procedures

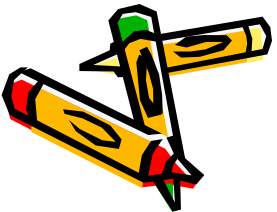
- Could only ask for Match to content specifications (collapse match and cognitive complexity)
- Could inform panelists of the intended match and ask for agree/verification (maybe as a follow up when mismatches are revealed)

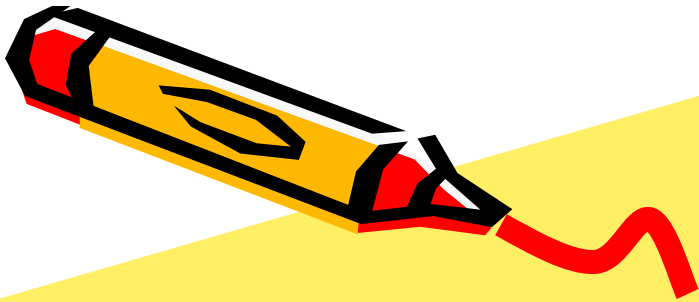


Importance of Alignment Studies



- Fundamental validity component
- If this piece isn't demonstrated, other test development, scoring, and interpretation steps are meaningless.
- No standard setting activity, no matter how well it is conducted, can make up for the lack of alignment in test scores to test framework.





Confessions of a Standard Setter

***Five Falsehoods that Foul
our Finest Efforts***

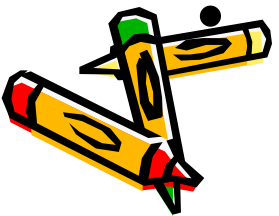
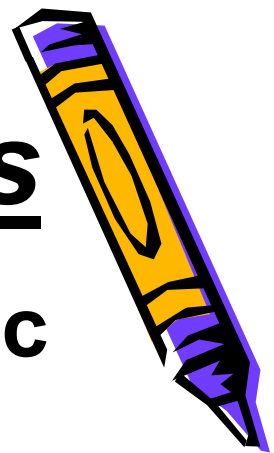
Mike Beck, BETA, Inc.



(HIGHLY CONFIDENTIAL: Do Not Share with “Believers”)

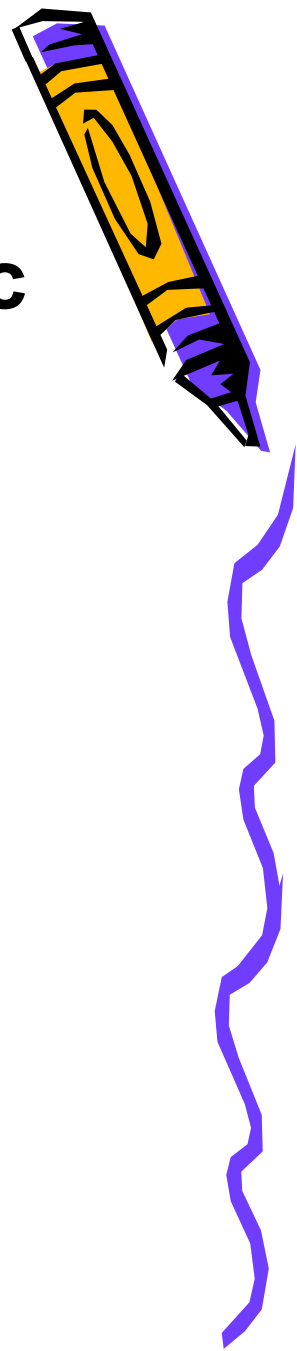
Five Fouling Falsehoods

- **Standards = Empirical, Scientific**
- **The method controls the variance**
- **You should try this at home**
- **Proficient = Proficient**
- **Pass = Pass**



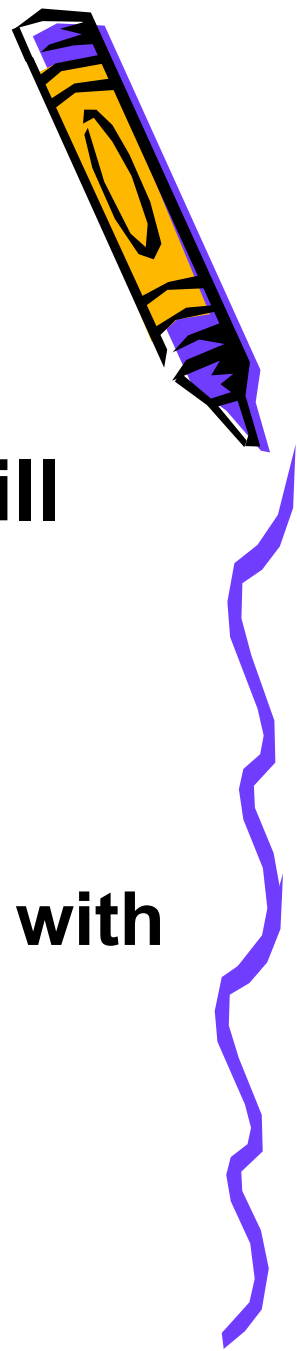
- **Standards are Empirical & Scientific**
 - Data do not define “science”
 - What’s wrong with “judgments”?

- **Method controls the variance**
 - Method *is* a variable, but . . .
 - It’s a distraction.
 - 56 other things are more important



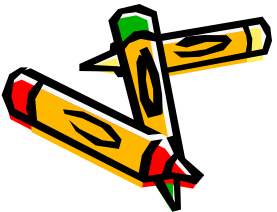
- **Don't try it yourself!**
 - *Anyone* can do this stuff
 - Two issues: “insider” & skill

- **Proficient = Proficient**
 - The terms control the results
 - We've nearly ended this mistake with *norms*, why not with *standards*?
 - Issues with NAEP as “truth”



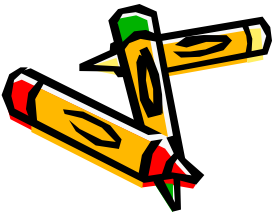
Definitions of “Proficient”

- **Solid** academic performance
- **Satisfactory**
- **Mastery** of grade-level standards
- **Solid** academic performance at grade level
- **Command** of challenging subject matter
- **High** level of achievement
- **Sound application** of higher-order skills
- **Acceptable** mastery of fundamental skills



- **Pass = Pass**

- **An issue of “retakes” and error**



False Positives & Negatives



<u>RS</u>	<u>State PR</u>	<u>Prob. of Passing</u>		
		<u>1 try</u>	<u>4 tries</u>	<u>8 tries</u>
34	25th	50%	74%	99.99%

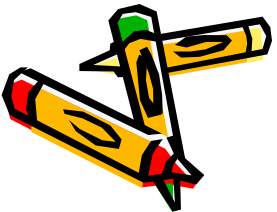
(50-item test; $SE_m = 3$. Assumes a few normal things, including no “growth” in the variable over time.)





Prob. of Passing

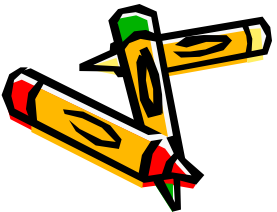
<u>RS</u>	<u>State PR</u>	<u>1 try</u>	<u>4 tries</u>	<u>8 tries</u>
36	31st	67%	85%	99.99%
34	25th	50%	74%	99.6%
31	18th	16%	45%	72%
29	13th	7%	26%	45%



Making Standards Consistent Across Grades / Content Areas



The whole issue:
“controlling variance”



Standardizing Standards Across Grades / Content Areas



- **Before you start:**
 - Consistency in the *content standards*
 - Consistency in *test difficulty* across tests

Issues in describing the labels:

Panel composition a key

Minimize Facilitator variance

Consistent, common starting point

(Luck)



How to “smooth”

(Once more, it’s not really a statistical issue:
“vertically integrated standards” BAH!)

Methods:

Statistical -- Finally, a *use* for scaled scores

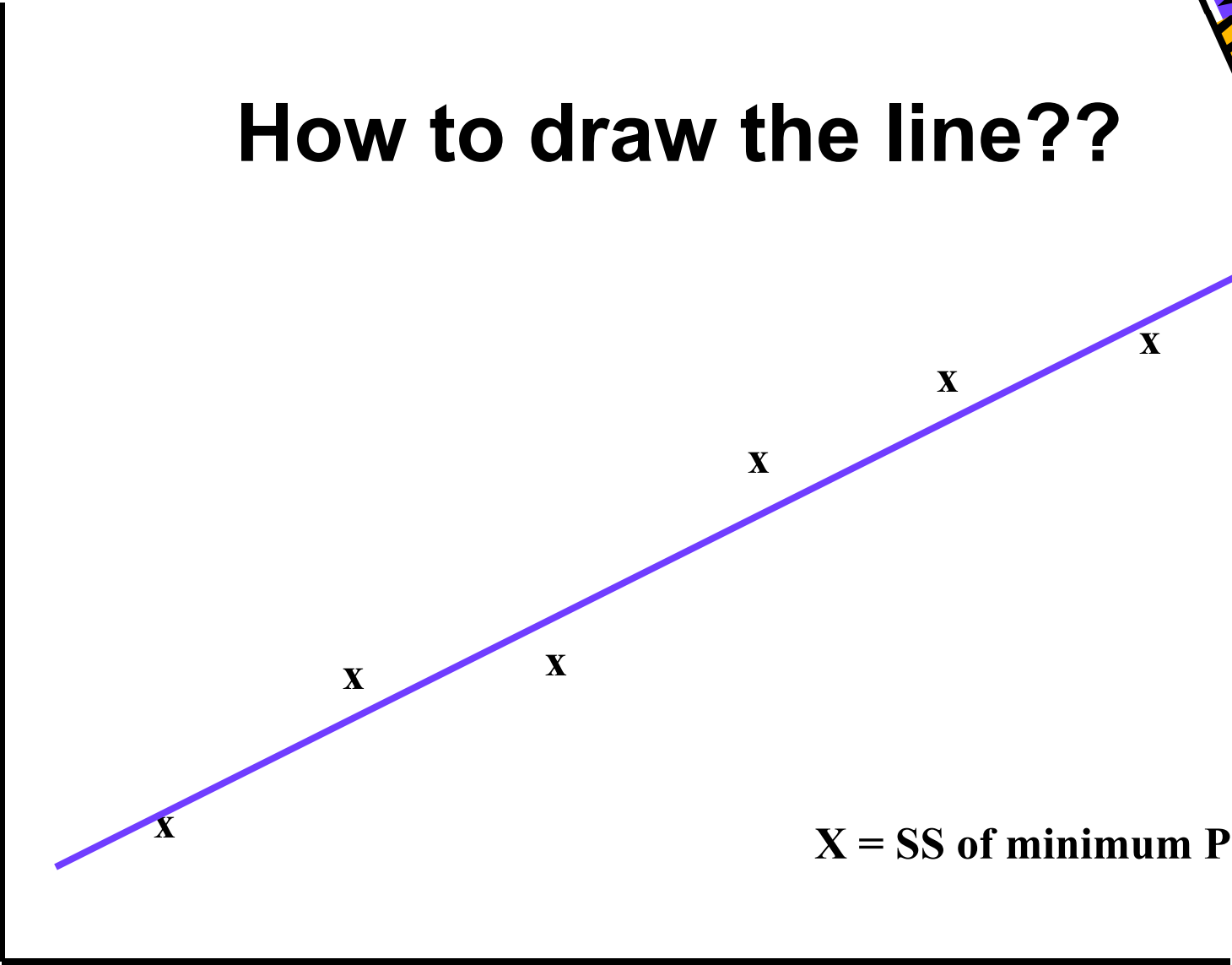
Judgmental -- Continue to rely on people

Mixed



How to draw the line??

S

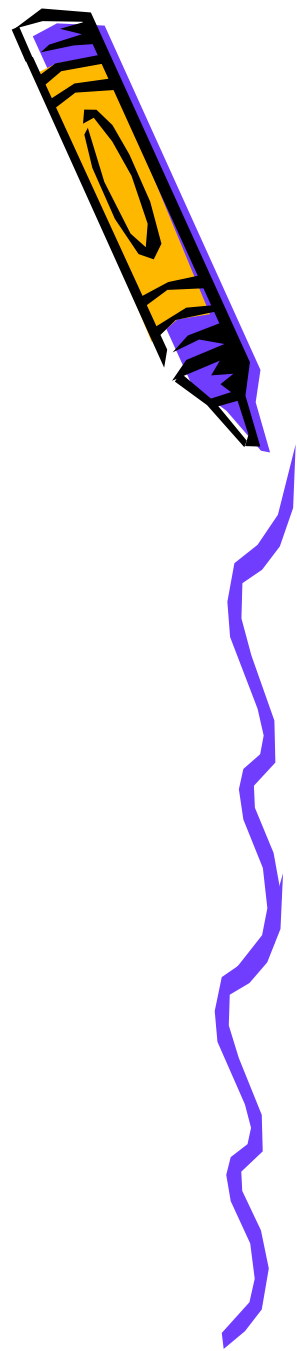


X = SS of minimum Pass

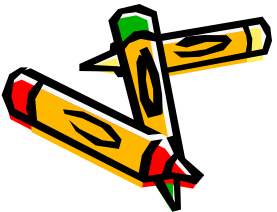
Grade Level



Other Issues in Cross-Test Standard Setting

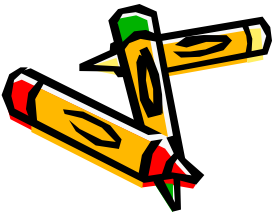
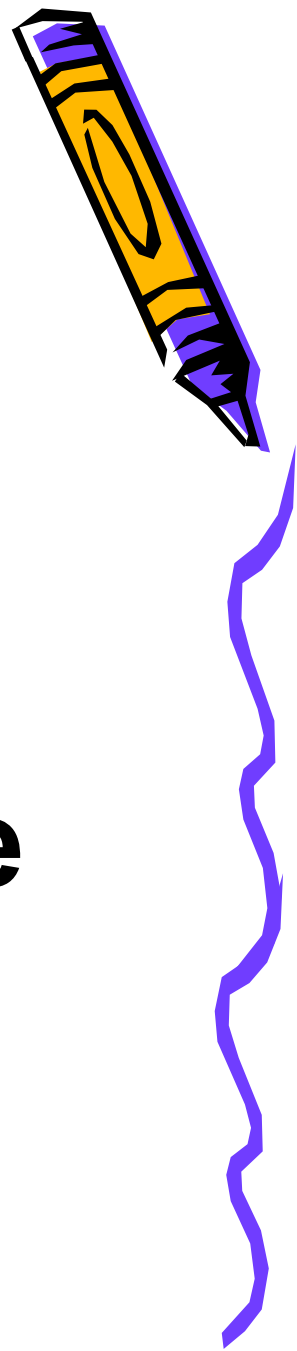


- *Must* consider the stakes -- panelists do.
- Only norms “equate” across grades and content areas.
- Don’t dismiss a possibility:
differences could be real.
- *No* way of adjusting is superior to not needing to adjust!



***What saves us . . .
and yet bothers us
most:***

- **You'll never know the truth.**





- The slides for the presentations in this session will be available on the publication page at the Center for Assessment website at www.nciea.org

