

Online Balanced Assessment System

CERA Presentation

12/1/11

Bill Conrad
David Foster
Mark Moulton

Presentation Outcomes

- Understand the key elements of the online performance assessment (OPS) system
- Learn how the OPS supports teacher professional learning communities
- Learn how the MARS math performance assessment system integrates into the OPS.
- Understand how an innovative scaling system for both performance and selected response assessments can be used formatively as well as a predictor of student performance on state tests

What Are the Problems?

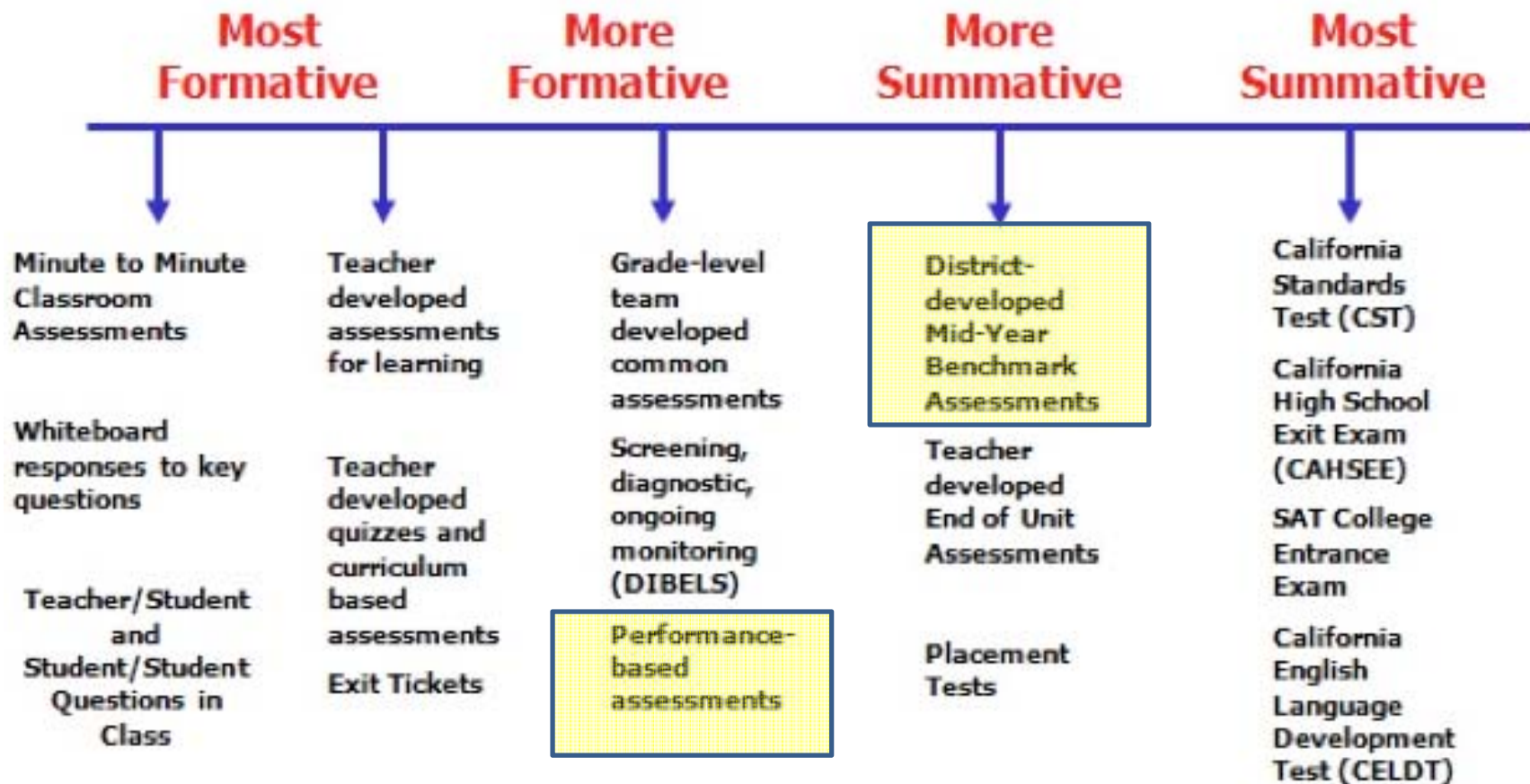
- How can school teams easily access performance assessments and resources aligned to the Common Core Standards?
- How can teachers easily scan and score MARS performance assessment tasks in an online environment?
- How can teachers easily share individual student results with students in ways that support student learning?
- How can teachers collaborate to calibrate their expectations for student performance as well as the quality of the assessment?

What is Our Vision?

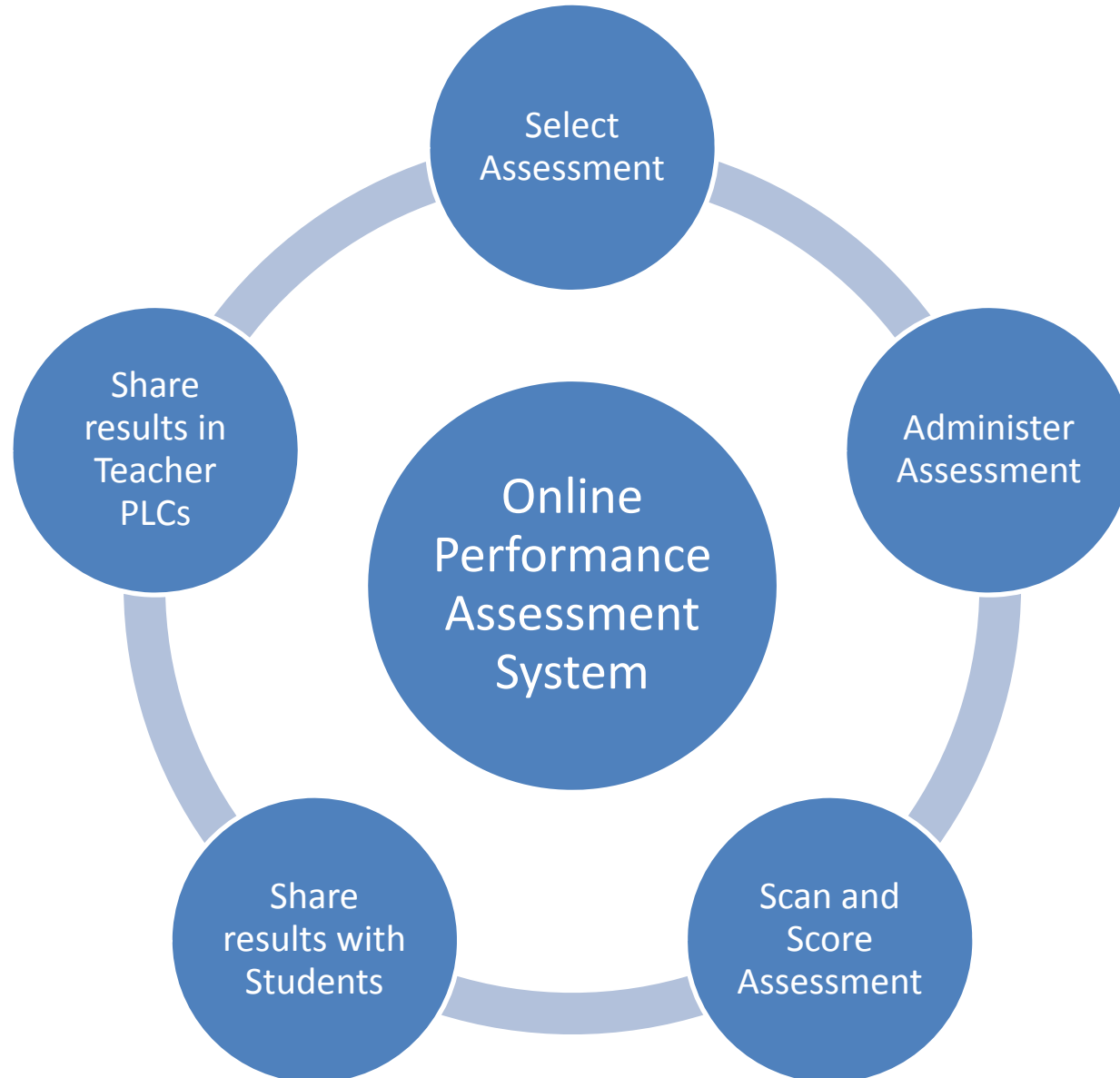
Build an online balanced assessment system that will support teacher teams in assessing student performance and thinking as students solve problems that are complex, real world and interesting.

Foster teacher team inquiry that informs the improvement of student learning.

WHERE ARE WE FOCUSED ON THE ASSESSMENT CONTINUUM?



System Overview



What is Our Implementation Timeline

Activity	Date
Collaborate on Building Prototype System	January, 2012
Volunteer Districts Pilot the Prototype System	February – May, 2012
Refinements Made to the System Based on the Pilot	June - July, 2012
Preparation for a Pilot of the Beta System in the Fall of 2012	August, 2012
Fall Pilot of the Beta System	September – November, 2012

What are Our Future Plans for the System?

- Create an online repository of online MARS Assessments and Resources that include performance tasks, rubrics, exemplars, error patterns and re-engagement lessons.
- Build the online Professional Learning element of the system to support teacher calibration and inquiry.
- Develop a scaling system that will support teachers in better using the system in a formative way.
- Develop systems to empirically test the validity and reliability of the assessments after they are administered.

Introduction to MARS Performance Assessments

Silicon Valley Mathematics Initiative

www.svmimac.org

Goals of Assessment

“We must ensure that tests measure what is of value, not just what is easy to test. If we want students to investigate, explore, and discover, assessment must not measure just mimicry mathematics.”

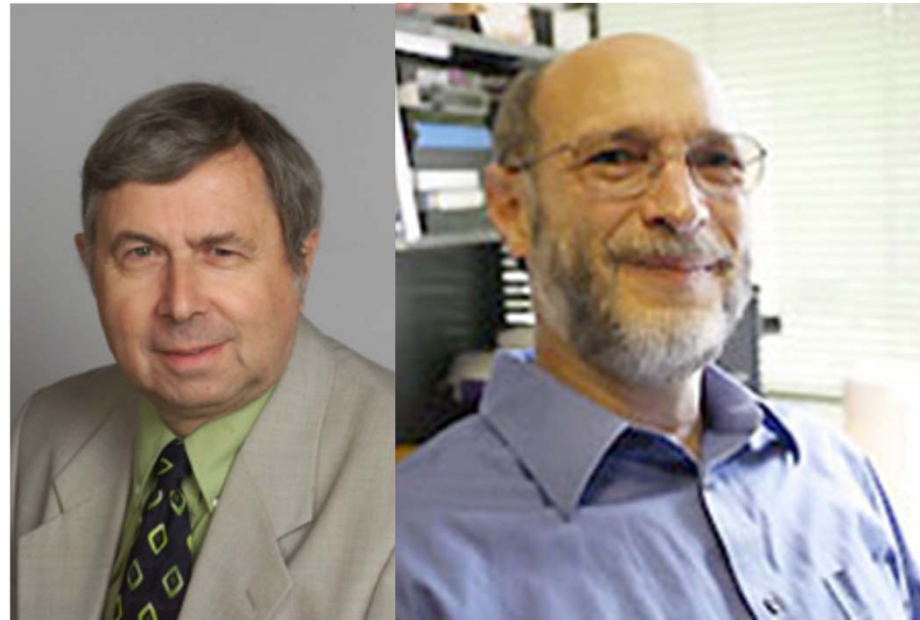


Everybody Counts

WYTIWYG

Gareth Mills at QCA

What you TEST is What you GET!



Next Generation Assessments – MARS Performance Tasks

Exemplars for both SBAC and PARCC Consortia

Performance Assessments


To Inform Instruction And Measure Higher Level Thinking

The Baker

This problem gives you the chance to:

- choose and perform number operations in a practical context

The baker uses boxes of different sizes to carry her goods.



Cookie boxes hold 12 cookies.
Donut boxes hold 4 donuts.
Muffin boxes hold 2 muffins.
Bagel boxes hold 6 bagels.

- On Monday she baked 24 of everything.
 How many boxes did she need? Fill in the empty spaces.
 cookie boxes _____ donut boxes _____
 muffin boxes _____ bagel boxes _____
- On Tuesday she baked just bagels. She filled 7 boxes.
 How many bagels did she make? _____
 Show your calculations.
- On Wednesday she baked 42 cookies.
 How many boxes did she fill? _____
 How many cookies were left over? _____
 Explain how you figured this out.

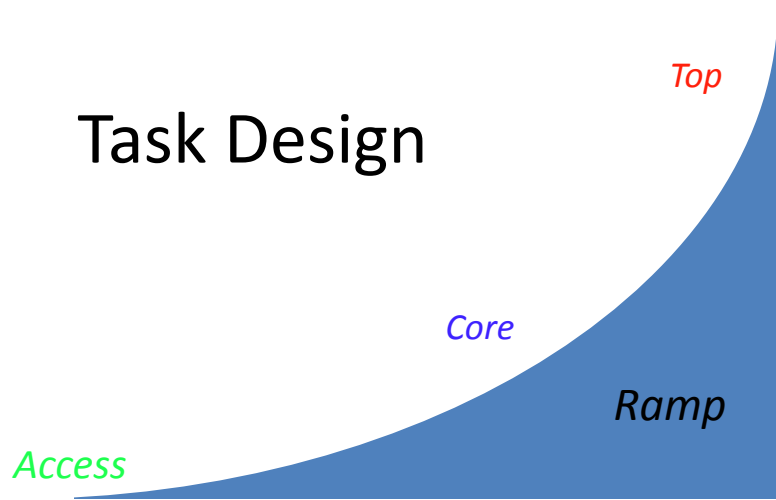
- On Thursday she baked 32 of just one item and she filled 8 boxes.
 What did she bake on Thursday? _____
 Show how you figured this out.

Copyright © 2007 by Mathematics Assessment Resource Service. All rights reserved.

Page 2

The Baker Test 4

Task Design



Entry level (access into task)

Core Mathematics - (meeting standards)

Top of Ramp (conceptually deeper, beyond)

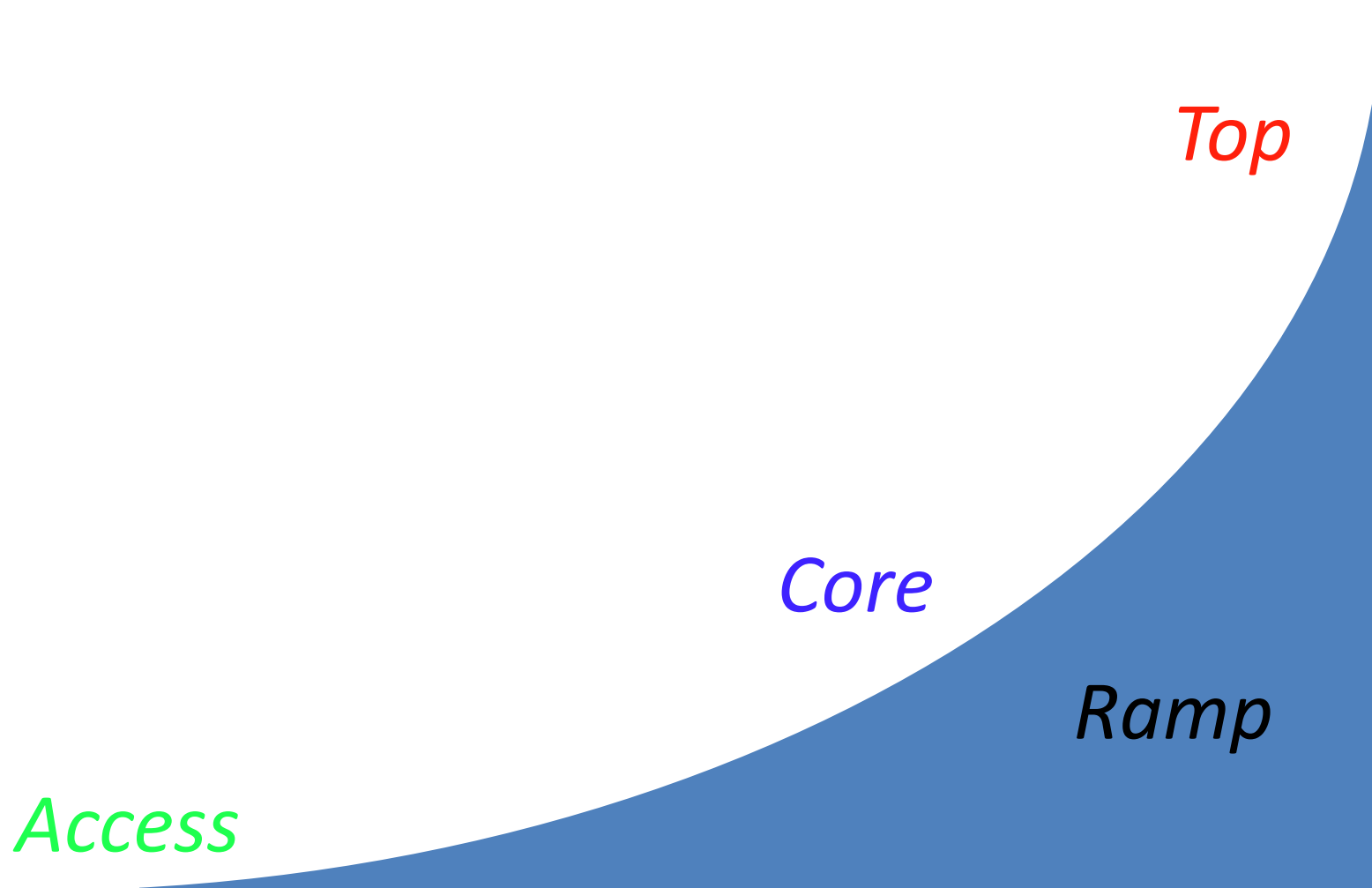
- The Mathematics Assessment Resource Service (MARS) is an NSF funded collaboration between U.C. Berkeley and the Shell Centre in Nottingham England.
- The Assessments target grades 2- Geometry and are aligned with the State and NCTM National Math Standards.



**BALANCED
ASSESSMENT**



The Design of a MARS Task



Dimensions of the Tasks

- Mathematical Content: CCSSM Domains
- Process Dimension: *Modeling and Formulating, Transforming and Manipulating, Inferring and Drawing conclusions, Checking and Evaluating, Reporting*
- Task Type: *Non-routine, design, plan, evaluate and make a recommendation, review and critique, representation of information, technical exercise, definition of concepts*
- Openness
- Reasoning Length Varies

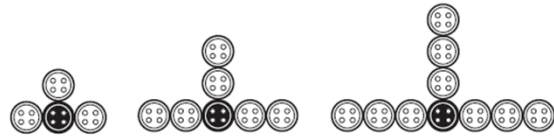


Buttons

This problem gives you the chance to:

- describe, extend, and make generalizations about a numeric pattern

Gita plays with her grandmother's collection of black and white buttons. She arranges them in patterns. Her first 3 patterns are shown below.



Pattern 1

Pattern 2

Pattern 3

Pattern 4

- Draw Pattern 4 next to Pattern 3.
- How many **white** buttons does Gita need for Pattern 5 and Pattern 6?

Pattern 5 _____

Pattern 6 _____

Explain how you figured this out.

- How many buttons in all does Gita need to make Pattern 11?

Explain how you figured this out.

- Gita thinks she needs 69 buttons in all to make Pattern 24.

How do you know that she is **not** correct?

How many buttons does she need to make Pattern 24?

Buttons		Test 5 Form A Rubric	
<p>The core elements of performance required by this task are:</p> <ul style="list-style-type: none"> describe, extend, and make generalizations about a numeric pattern <p>Based on these, credit for specific aspects of performance should be assigned as follows:</p>		Points	Section Points
1.	Draws correct Pattern 4.	1	1
2.	<p>Gives correct answers as:</p> <p>Pattern 5 = 15 white buttons Pattern 6 = 18 white buttons</p> <p>Gives explanation such as:</p> <p>Pattern 5 has 3 rows of 5 white buttons = 15 white buttons and Pattern 6 has 3 rows of 6 white buttons = 18 white buttons.</p> <p><i>Special case:</i> Gives answers 16 and 19 with correct explanations including black buttons.</p>	1 1 1 2 s c	 3
3.	<p>Gives correct answer as:</p> <p>34</p> <p>Gives explanation such as:</p> <p>Pattern 11 has 3 rows of 11 white buttons = 33 white buttons and 1 black button.</p> <p><i>Special case:</i> Gives answer as 33 with a correct explanation for the white buttons.</p>	1 1 1 s c	 2
4.	<p>Gives explanation such as:</p> <p>Pattern 24 needs 3 rows of 24 white buttons = 72 white buttons and 1 black button, 73 buttons in all.</p> <p><i>Accept alternative correct explanations such as: 69 is divisible by 3, so it cannot be correct.</i></p> <p>Gives correct answer as:</p> <p>73</p>	1 1	 2
Total Points			8

Scoring Protocols

Each MARS Task is accompanied with:

- A specified Rubric
- Five Training Papers
- Ten Standardizing Papers
- A set of Scoring Protocols

Scoring Marks

- ✓ correct answer or comment
- x incorrect answer or comment
- ✓ft correct answer based upon previous incorrect answer called a follow through
- ^ correct but incomplete work - no credit
- () points awarded for partial credit.
- m.r. student misread the item. Must not lower the demands of the task -1 deduction

The Party

1. Darren and Cindy are planning a party for their friends. They have 9 friends coming to the party. How many people will be at the party? _____.
2. They are buying cupcakes and cans of soda. Cupcakes cost \$1.50 and soda costs 75¢. How much does it cost for each person? _____.
Show how you figured it out.
3. How much will it cost for everyone to have a cupcake and soda?

Show how you figured it out.
4. They just remembered to buy a 50¢ party bag for each friend. Show how to find the total cost for the party.

The Party

1. Darren and Cindy are planning a party for their friends. They have 9 friends coming to the party. How many people will be at the party? _____.
2. They are buying cupcakes and cans of soda. Cupcakes cost \$1.50 and soda costs 75¢. How much does it cost per person? _____. Show how you figured it out.
3. How much will it cost for everyone at the party to have a cupcake and soda? _____. Show how you figured it out.
4. They just remembered to buy a 50¢ party bag for everyone at the party. Show how to find the total cost for the party.

The Party -	Pts	Sec tio n
1. 11 people	1	1
2. \$2.25 Shows work such as: $\$1.50 + 75¢$	1 1	 2
3. \$24.75 Shows work such as: $11 \cdot \$2.25$	1 f.t. 2	 3
4. Shows work such as: $11 \cdot 50¢ = \$5.50$ $\$5.50 + \$24.75 = \$30.25$ <i>partial credit</i> only shows $11 \cdot 50¢$	2 (1)	 2
Total Points		8

The Party

1. Darren and Cindy are planning a party for their friends. They have 9 friends coming to the party. How many people will be at the party? 11

2. They are buying cupcakes and cans of soda. Cupcakes cost \$1.50 and soda costs 75¢. How much does it cost per person? \$2.50 ~~x~~ Show how you figured it out.

$$\$1.50 + 75¢ = \$2.50$$

3. How much will it cost for everyone at the party to have a cupcake and soda? \$27.50 ~~x~~ Show how you figured it out. ~~x~~

$$11 \cdot \$2.50$$

4. They just remembered to buy a 50¢ party bag for everyone at the party. Show how to find the total cost for the party.

$$11 \cdot 50¢ = \$5.50$$

x

1
0
1
1 ft
2
(1)
6

The Party -	Pts	Section
1. 11 people	1	1
2. \$2.25 Shows work such as: \$1.50 + 75¢	1	2
3. \$24.75 Shows work such as: 11 • \$2.25	1 f.t. 2	3
4. Shows work such as: 11 • 50¢ = \$5.50 \$5.50 + \$24.75 = \$30.25 <i>partial credit</i> only shows 11 • 50¢	2 (1)	2
Total Points		8

Scoring Process



- 1) Work the task yourself
- 2) Whole Group - go over the point scoring rubric
- 3) Individually score the five Training papers (T1 - T5)
- 4) Whole group - Review standard scores for T1 - T5
- 5) Individually score the 10 Standardizing papers
- 6) Whole group - Review standard scores for S1 - S10
- 7) Ready for “live” papers

Scoring Marks

- ✓ correct answer or comment
- x incorrect answer or comment
- ✓ft correct answer based upon previous incorrect answer called a follow through
- ^ correct but incomplete work - no credit
- () points awarded for partial credit.
- m.r. student misread the item. Must not lower the demands of the task -1 deduction

Performance Exams

40,000 – 70,000 students per year since 1999



Students in grades 2 through 10th/11th grade are administered performance exams (5 apprentice tasks per exam).

Task 1: Candies	Rubric	
	points	section points
The core elements of performance required by this task are: • work with fractions and ratios		
Based on these, credit for specific aspects of performance should be assigned as follows		
1. Gives correct answer: 2/3 or 6/9	1	1
2. Gives correct answer: 3 Shows work such as: $1 + 3 = 4$ $12 \div 4 =$ Accept diagrams.	1	2
3. Gives correct answer: 18 Shows work such as: $2 + 3 = 5$ $30 + 5 = 6$ $6 \times 3 =$ Accept diagrams.	2	3
4. Gives correct answer: 6 Gives a correct explanation such as: Anthony mixes a ratio of one cup of cream to two cups of chocolate. The ratio stays the same for different amounts. So I wrote the numbers in a chart like this: 1 to $2 =$ a total of 3 2 to $4 =$ a total of 6 3 to $6 =$ a total of 9 Accept diagrams.	1	3
Total Points		8

District scoring leaders are trained in using task specific rubrics



Student results are collected, analyzed, and reported by an independent data contractor.



Random sample of student papers are audited and rescored by SJSU math & CS students. (Two reader correlation >0.95)



Student tests are hand scored by classroom teachers trained and calibrated using standard protocols.

MARS vs. CST

Silicon Valley Mathematics Initiative

MAC Final Data Spring 2011

MARS Exam Spring 2011

Grade Level or Course	Number Students Assessed
Second Grade	6585
Third Grade	5779
Fourth Grade	6005
Fifth Grade	7183
Sixth Grade	5142
Seventh Grade	3719
Eighth Grade	755
Course 1 (Algebra 1)	2938
Course 2 (Geometry)	432

A total of 38,538 students were administered MARS tests during the spring 2011. That includes 9 grade/course levels, 28 districts from six counties in the great San Francisco Bay Area.

Spring 2011 Trends Grade to Grade

Grade 2	MARS 1	MARS 2	MARS 3	MARS 4	Total
Far Below	1.0%	0.6%	0.1%	0.0%	1.7%
Below Basic	1.9%	4.1%	1.1%	0.1%	7.2%
Basic	0.8%	5.3%	4.6%	0.6%	11.3%
Proficient	0.4%	5.1%	16.2%	6.5%	28.2%
Advanced	0.2%	1.6%	15.2%	34.6%	51.6%
Total	4.3%	16.7%	37.2%	41.8%	100.0%

Grade 2	MARS Below	MARS At or ^	Total
CST Below	13.7%	6.5%	20.2%
CST AT or ^	7.3%	72.5%	79.8%
Totals	21.0%	79.0%	100.0%

Spring 2011 Trends Grade to Grade

Grade 3	MARS Below	MARS At or ^	Total
CST Below	16.4%	4.5%	20.9%
CST AT or ^	12.7%	66.3%	79.0%
Totals	29.1%	70.8%	99.9%

Grade 4	MARS Below	MARS At or ^	Total
CST Below	15.6%	5.8%	21.4%
CST AT or ^	12.9%	65.8%	78.7%
Totals	28.5%	71.6%	100.1%

Grade 5	MARS Below	MARS At or ^	Total
CST Below	17.3%	6.0%	23.3%
CST AT or ^	12.4%	64.4%	76.8%
Totals	29.7%	70.4%	100.1%

Grade 6	MARS Below	MARS At or ^	Total
CST Below	34.7%	3.8%	38.5%
CST AT or ^	21.7%	39.6%	61.3%
Totals	56.4%	43.4%	99.8%

Spring 2011 Trends Grade to Grade

Grade 7	MARS Below	MARS At or ^	Total
CST Below	38.1%	0.4%	38.5%
CST AT or ^	38.1%	23.5%	61.6%
Totals	76.2%	23.9%	100.1%

Grade 8	MARS Below	MARS At or ^	Total
CST Below	55.1%	2.8%	57.9%
CST AT or ^	25.0%	17.0%	42.0%
Totals	80.1%	19.8%	99.9%

Course 1	MARS Below	MARS At or ^	Total
CST Below	31.9%	4.1%	36.0%
CST AT or ^	21.5%	42.0%	63.5%
Totals	53.4%	46.1%	99.5%

Course 2	MARS Below	MARS At or ^	Total
CST Below	15.4%	0.0%	15.4%
CST AT or ^	36.0%	48.7%	84.7%
Totals	51.4%	48.7%	100.1%

8th Grade Geometry

California's Highest Achieving Students

Geometry	MARS Below	MARS At or Above	Total
CST Below	15.3%	0.0%	15.3%
CST AT or Above	36.0%	48.7%	84.7%
Totals	51.3%	48.7%	100%

Link Assessment and Learning



“Assessment should be an integral part of teaching. It is the mechanism whereby teachers can learn how students think about mathematics as well as what students are able to accomplish.”

Everybody Counts

EDS Provides the Technical Platform

- Online repository of Common Core aligned items and tests.
 - MC, PT, or a mix (e.g., MARS)
- The Teacher/School/District downloads and prints a test, including pre-ID answer sheets.
- Scanning solution – any scanner will work so long as it meets EDS software requirements.
- Scanned data goes straight to EDS servers.
 - MC items get scored
 - PT item responses are saved as images. Designated raters log in and score the images. They also add feedback for student consumption.
- Scores/student feedback/reports are accessible online.
- Data goes to professional learning community and raters.

What I'm Going to Talk About

- Scaling. EDS has been developing a suite of psychometric products under the name EdScale.
 - Goal: Convert MC/PT scores from any test into a comparable standard metric, such as CST or Common Core scale scores.
- To be *really* useful, online performance assessment systems will need this capacity to:
 - Compare to State/Federal definitions of proficiency (AYP, etc.)
 - Measure student growth across the year
 - Estimate value-added
- My topic: “How to Have Our Cake and Eat it, Too”

Formative vs. Summative

- The “formative” need
 - Teachers need quick, actionable student-level information in order to adapt instruction daily and weekly
 - Requires locally written/selected items, especially PT
- The “summative” need
 - Districts, principals, teachers need where students are relative to each other and to state/federal standards (e.g., AYP).
 - Increasingly need multiple data points for growth measures (value-added measures)
 - Requires tests aligned to state/federal standards, generally MC, that meet psychometric standards of reliability
- The universal need: more time on education, less on testing
- Problem: Good formative tests make bad summative tests. Good summative tests make bad formative tests.

Solution

- Make each benchmark and formative exam ***double*** as a summative test. Make it psychometrically equivalent to a CST or Common Core test:
 - Yet have tests/items be locally responsive, aligned to recent instructional content
 - Without requiring that all items be specifically aligned to state or federal standards
 - With flexibility to remove under-performing items and continually adapt and improve tests

Some Psychometric Issues

- How to equate local tests to each other and to standard (state or common core) tests.
 - Issue 1. Successive local tests may not share a common construct or that of the standard test.
 - Issue 2. Cut-scores assigned to local tests may poorly match those of the standard test.
 - Issue 3. A regression equation relating local scores to standard scores quickly goes out of date.
 - Issue 4. PT raw scores can be unreliable.
 - Issue 5. It is hard to analyze PT items and MC items together as they tend to form at least two dimensions.
 - Issue 6. It is hard to track common items across tests for equating. Nor can local tests be linked by common items to state or common core tests.

Multidimensional IRT

- EDS addresses these issues with an alternating least squares multidimensional matrix decomposition algorithm called NOUS.
 - MC items (at the distractor level), PT items, and standard scale scores (e.g., CSTs,...) are modeled as vectors in a common mathematical space, generally of 2 or 3 dimensions. These vectors are built by ***transferring information between the columns*** (and rows).
 - Common Item Equating. When PT and MC items are given around the same time as the CSTs, they can be calibrated together. Any subset of items can then be used to estimate likely performance on the CST dimension on future tests.
 - Common History Equating. When common item equating is not feasible, equating is done using scores from a test (CST) administered to each student in the previous year. Local item performance is converted into performance on last-year's CST dimension, which is aligned to the current-year CST dimension (Math, ELA). This is combined with overall annual student growth rates to estimate expected performance on the current year's CST.

Observed Responses, Mixed PT and MC

Student	Grade	Task 1	Task 2	Task 3	Task 4	Task 5	MC 1	MC 2	MC 3	MC 4	MC 5	Math CST, Actual
1	5	3	0	2	2	0	C	B	C	C	B	295
2	5	3	1	3	3	0	B	B	D	A	B	295
3	5	4	1	3	3	0	C	B	B	D	B	285
4	5	7	3	7	6			B	C	A	D	365
5	5	6	2	5	4			A	D	D	B	376
6	5	8	4	7	7	1	D	C	D	D	D	414
7	5	3	1	3	2	0	D	B	B	C	C	280
8	5	2	1	2	2	1	D	A	D	C		295
9	5	1	1	1	1	0	D	A	B	A	A	295
10	5	6	1	4	4	0	D	D	D	A	C	319
11	5	5	4	4	4	3	A	D	B	C	D	407
12	5	6	2	6	5			D	C	B	B	407
13	5	1	1	1	1			A	C	A	A	344
14	5	2	1	1	1	0	C	D	D	D	A	260
15	5	5	1	4	3	0	A	B	C	D	D	289
16	5	8	6	8	8	3	B	C	A	D	B	495
17	5	7	3	6	6	1	C	C	D	D	A	344
18	5	8	3	8	7	0	C	C	C	C	A	422
19	5	7	1	6	6	0	A	D	B	B	C	319
20	5	4	1	3	3	0	A	C	B	B	C	324

From MARS 2011, Grade 5 (7183 students)							x	-258.5	-45.2	-212.2	-181.4	82.8	-42.3
							y	38.8	107.4	54.0	58.8	184.7	127.3

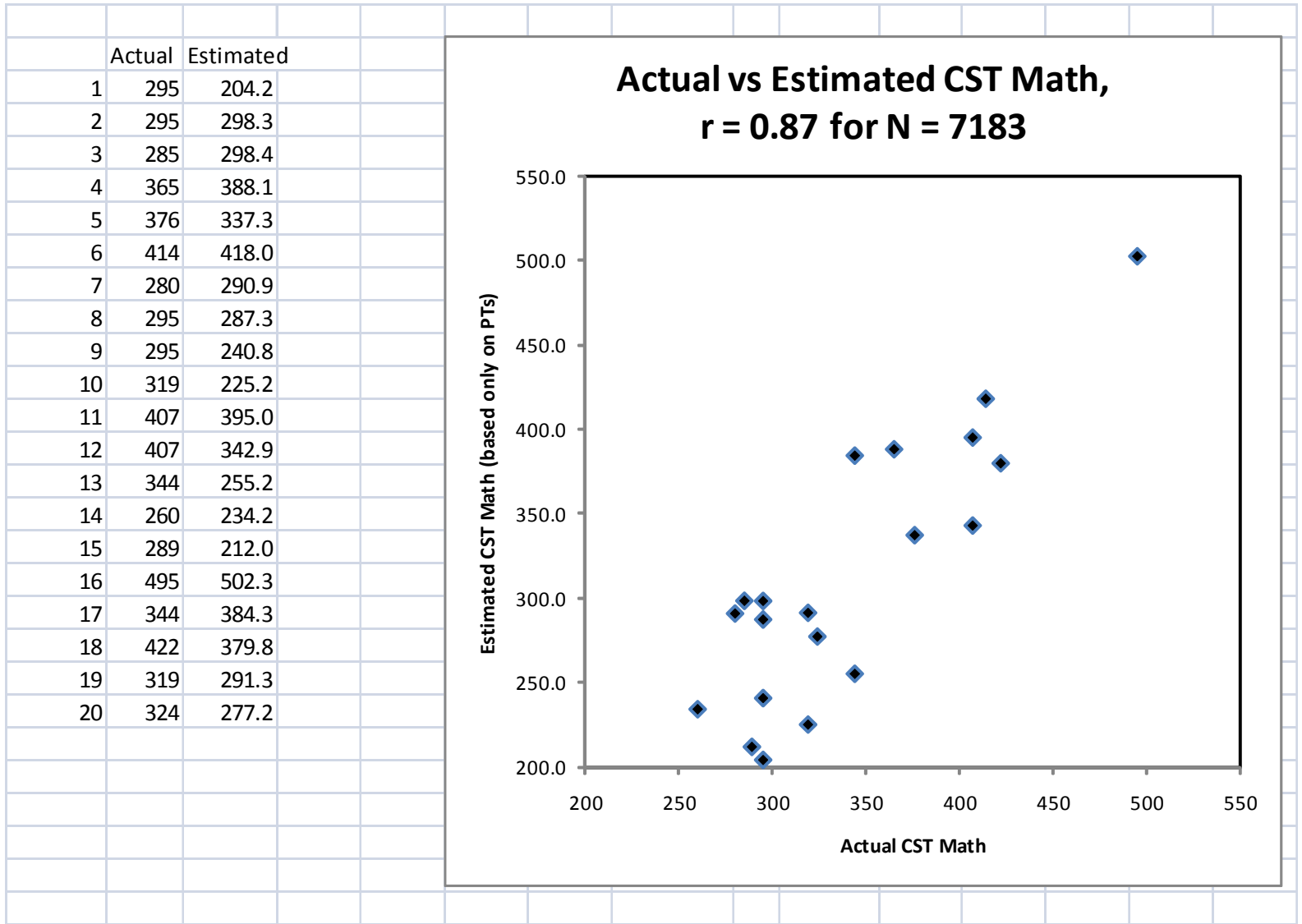
Observed Responses

Estimates (based on other data values)

Student	Grade	Task 1	Task 2	Task 3	Task 4	Task 5	Math CST, Actual	Student	Grade	Task 1	Task 2	Task 3	Task 4	Task 5	Math CST, Estimated from Tasks
1	5	4	0	5	1	0	295	1	5	3	0	2	2	0	204.2
2	5	4	1	3	3	1	295	2	5	3	1	3	3	0	298.3
3	5	6	1	3	4	1	285	3	5	4	1	3	3	0	298.4
4	5	7	4	7	7	1	365	4	5	7	3	7	6	1	388.1
5	5	6	1	6	4	1	376	5	5	6	2	5	4	1	337.3
6	5	8	5	7	6	1	414	6	5	8	4	7	7	1	418.0
7	5	5	1	3	2	1	280	7	5	3	1	3	2	0	290.9
8	5	0	3	3	2	0	295	8	5	2	1	2	2	1	287.3
9	5	6	0	0	3	2	295	9	5	1	1	1	1	0	240.8
10	5	6	0	7	3	0	319	10	5	6	1	4	4	0	225.2
11	5	4	5	5	2	1	407	11	5	5	4	4	4	3	395.0
12	5	6	3	5	3	0	407	12	5	6	2	6	5	1	342.9
13	5	6	0	4	0	2	344	13	5	1	1	1	1	0	255.2
14	5	6	1	2	0	0	260	14	5	2	1	1	1	0	234.2
15	5	6	0	5	4	0	289	15	5	5	1	4	3	0	212.0
16	5	8	5	8	8	5	495	16	5	8	6	8	8	3	502.3
17	5	6	5	7	5	1	344	17	5	7	3	6	6	1	384.3
18	5	8	4	7	7	0	422	18	5	8	3	8	7	0	379.8
19	5	4	1	8	4	0	319	19	5	7	1	6	6	0	291.3
20	5	6	0	4	5	3	324	20	5	4	1	3	3	0	277.2

* In "common history" design, Math CST is from previous year.

In "common item" design, Math CST is current year and items are calibrated and banked for future use.



MARS 2011, Grade 5

Summary PT/CST Stats

MARS, Grade 5 Performance Tasks	Mean	SD	Corr	Resid	RMSE	Rel	Fit_MeanSq	Fit_Perc>2	Count	Min	Max
MARS Task 1	6.8	1.8	0.78	1.2	1.46	0.35	4.4	0.14	7183	0	8
MARS Task 2	4.1	2.1	0.82	1.3	1.71	0.35	1.6	0.11	7183	0	8
MARS Task 3	6.4	1.9	0.82	1.3	1.60	0.30	3.7	0.15	7183	0	8
MARS Task 4	6.2	1.9	0.79	1.3	1.63	0.28	2.8	0.15	7183	0	8
MARS Task 5	2.2	2.3	0.83	1.3	1.66	0.46	2.4	0.13	7183	0	8
Math CST Estimates, from Tasks	413.9	100.0	0.87	50.7	67.10	0.55	1.2	0.05	7145	74.5	682.2

Snapshot Item Analysis (2-dimensional solution was clearly optimal)

The PT-based Math CST estimates correlate well (0.87) with the actuals. This compares well with many much longer MC tests. (Actually, MARS is more like a 40 item than a 5 item test.)

Reliability seems low (0.55) due to high RMSE (67.1). The RMSE is driven by the small number of “items” (5) and the residuals between the CST estimates and actuals. If the CST were being compared to itself instead of the MARS, the RMSE would be around 31.1 ($= \sqrt{22^2 + 22^2}$), where 22 is the published Math CST Grade 5 measurement error. The reliability of the individual MARS tasks is relatively low, in part because students tend to clump around the mean. The standard deviations (SD) and reliability could be increased with some judicious rewriting around the task steps near the mean.

The fit statistics for the CST estimates are optimal – we would expect 5% of scores to misfit by chance (be significantly different from the actuals), and that is what we see. However, the misfits for the performance tasks are a bit high. About 15% of students give significantly unexpected answers, when we would expect 5%. Lots of factors can cause students to misfit the test, including confusions with the item.

Overall, the test seems fairly strong given its apparent small size, and it is a suitable proxy for the CST.

How Multidimensional Helps

- Issue 1 – Common Construct. By using each test only *insofar as* it predicts a common scale (e.g., CSTs), we make the scale scores comparable across tests. Now we can measure growth on a commonly understood metric.
- Issue 2 – Cut-Scores. By predicting CST scores, we automatically have access to California (or federal) definitions “proficiency”. No arbitrary cut-scores set locally.
- Issue 3 – Out of Date Regression Equations.
 - Common Item Equating. Since prediction is done at the level of the individual item, tests can be changed as long as each contains items from a pre-calibrated bank.
 - Common History Equating. The dependent variable is past CST performance, not future performance, so we don’t need an already existing regression equation to predict future CST performance. That gets handled using last year’s scores and estimating average cross-grade student growth for the sample.

How it Helps (cont.)

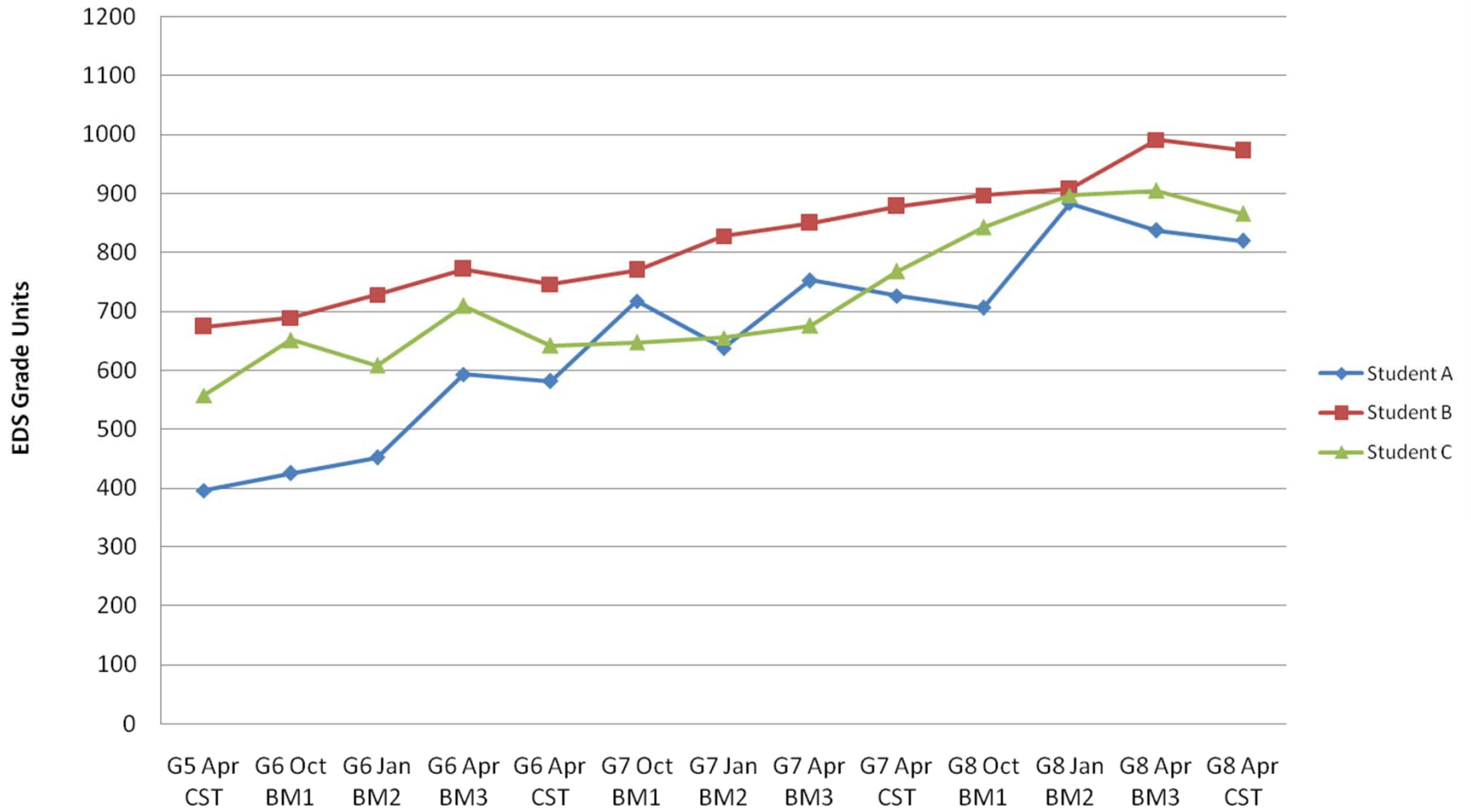
- Issue 4 – PT Scores Unreliable. Actually, well-built PT items are often more reliable than MC items, even with human scoring. The problem is, there's not enough of them. Our algorithm mitigates this problem by trading information across all items (MC and PT), improving the reliability of each.
- Issue 5 – PT and MC are Different Dimensions. Use of a multidimensional model largely addresses this problem. To a large extent, they can be modeled in a common 2- or 3- dimensional space.
- Issue 6 – Common Item Designs are a Pain. This is addressed using the common history equating model which requires no linking items.
- What this Means. Districts, schools and teachers can target local assessments for formative purposes, yet realize the benefits of State and Common Core summative testing. Plus, you get **growth**.

Some Ways to Use Scaling Results

- AYP. Use formative tests to forecast AYP and other statistics. Get a sense for how the kids are doing relative to the rest of the state
- Evaluation. One data point per year (the CST) is completely inadequate for evaluating teacher value-added. We need at least five. Scaled formatives can fill in the gap.
- Test Improvement. Every test receives a complete item analysis which teachers and the professional learning community can use to improve it.
- Diagnostics. Separate scores can be reported for each standard, or by item type. Due to inter-item information trading, they are more reliable than traditional raw scores.
- Growth. Within-year and cross-year student growth is intrinsically interesting – we should do more of it.

Three "Students," Grades 6 - 8*

(*Student trend-lines are rank-matched across grades to create growth examples.)



EdScale Status

- EdScale has been at use in a variety of school districts, including SCCOE, for around four years. Our research (see www.eddata.com) shows that it is performing well. There are some issues with upper-grade Math.
- EdScale has supported only MC items, one equating model, and two contents (Math and ELA). The next version will support all item types, such as PT, at least five different equating models, and a variety of content areas.
- Our goal is to have the new version of EdScale available for use formative and benchmark testing programs before the next school year.

To Have Our Cake and Eat It, Too

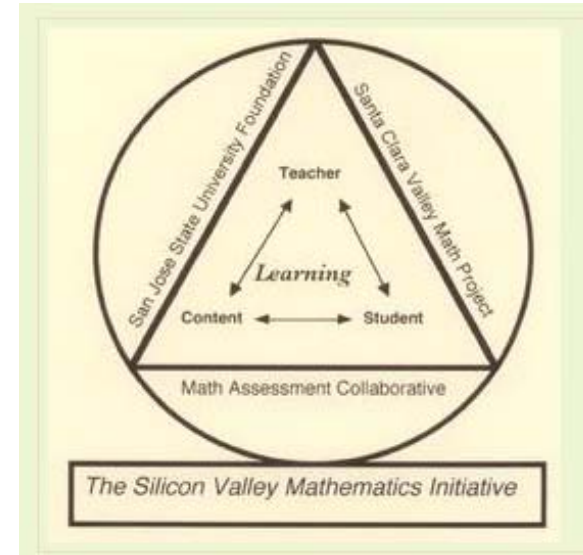
- Pool. Pool local educational expertise (PLC) to create smart, well-targeted, educationally meaningful tests.
- PTs. Use lots of performance tasks. Take advantage of the online scoring process to see student work directly.
- Automate. Automate the process of maintaining items and tests, scanning, scoring, scaling and reporting. Make it easy to give kids feedback. Focus on education, not testing.
- Mini-CSTs. Meanwhile, without shifting focus from education, use scaling to get CST-equivalent or Common Core-equivalent scale scores using locally built and selected formative exams (and some psychometric magic).
- Growth. Compare students with each other and across time, even as the tests change.
- AYP. Forecast percent “proficient” on the standard metric – no need for standard setting to set cut-points.
- Improve Tests. Use item statistics and tools such as “Wright maps” to iteratively increase the psychometric power of formative tests.
- PLCs. Build the professional learning community around control of inputs and outputs and pride of craft.

Presenters

Bill Conrad
Santa Clara County Office of Education
408-453-4332 (Office)
510-761-2007 (Cell)
Bill_Conrad@sccoe.org

David Foster
Silicon Valley Math Initiative
(408) 776-1645 (Office)
(408) 472-5706 (Cell)
dfoster@svmimac.org

Mark Moulton
Educational Data Systems
(408) 776-7646 (Office)
(408) 710-5197 (Mobile)
markhmoulton@gmail.com



Online Balanced Assessment System

CERA Presentation

12/1/11

Bill Conrad
David Foster
Mark Moulton