What do we know now?
Reflections on the Spring 2012 Mathematics Pilot Test Results

Alicia Stoltenberg
Perneet Sood
Balanced Systems of Assessment

- **Formative** - used frequently to provide additional information about student mastery of specific standards
- **Interim** - given up to 3 times a year to provide information about student needs in order to adjust instructional programs
- **Summative** - administered once a year to provide information for state and federal evaluation of the progress of Kansas students

Formative – **process** not a test

Interim – “benchmark-like”
- covers the whole curriculum – for a predictive score
- benchmark tests are probably closer to quarterly tests

Summative – KS Assessment, NCLB Assessment
- what words do you associate with this test?
Balanced Systems of Assessment

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Context</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monitor and improve learning</td>
<td>Classroom - teachers, students</td>
<td>Formative</td>
</tr>
<tr>
<td>Evaluate instructional programs</td>
<td>Grade-level teacher teams/PLCs</td>
<td>Interim</td>
</tr>
<tr>
<td>Inform policy/judge institutional effectiveness</td>
<td>Building/district</td>
<td>Summative</td>
</tr>
</tbody>
</table>

Formative—does not have to be a formal “test”
Our interim assessments don’t provide instructionally sensitive feedback. From an assessment perspective, the tests are too short to provide solid feedback about what a student knows (especially at the middle achievement level) Good information about top and bottom achievement levels.

Discussion opportunity:

How do you use the different tiers of assessment in your classroom, department, and/or school?
  In other words, what do these tiers of assessment look like for you?

What types of formative assessment do you use that are not tests?
Brookhart suggests that the lower left quadrant should be the biggest piece of the pie. In other words, if we are viewing the current format as four equal parts, we really should move the origin up closer to the top right, so that the lower left quadrant should be the largest part.

It is dangerous to use data in a way that it was not intended to be used—only use the data to make decisions for the reasons it was originally collected!
For more information on CCSSM progressions, do a Google search for “math progressions”

Site address: http://ime.math.arizona.edu/progressions/#products
The arrow is larger to the right on the bottom because, while this is a double-headed arrow, instruction should influence assessment MORE than assessment influences instruction.

Of course assessment results can be used to inform instructional decisions, but be careful to use results in the way they were intended. Using results in ways other than how they were intended can have poor consequences.
CETE developed the interim and formative that are currently available. We only administer the summative.

CETE does not have any affiliation with SBAC, so we only know as much as everyone else about what the tests will look like.
Test Development Process

**Statement**
- Content to be assessed
- Purpose of assessment

**Specifications**
- Test
- Individual items

**Item Development**
- Meet the needs of the assessment
- Internal review

**External Review**
- Content review
- Bias review

**Item Selection**
- Item content
- Statistical properties

**Field Testing**
- Statistics gathered

**Administration**
- Verify scaling and equating

**Equating**
- Achievement standards have consistent meaning

**Quality Checks**
- Item content
- Statistical properties

**Reporting**
- Statistics gathered
# 2012 Embedded Standards

<table>
<thead>
<tr>
<th>Elementary Level</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>- Understand fractions as numbers</td>
<td></td>
</tr>
<tr>
<td>- Use/create multiple representations of fractions</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>- Extend understanding of fractions as numbers</td>
<td></td>
</tr>
<tr>
<td>- Use fractions to solve problems, including multiplication of fractions</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>- Extend understanding of fractions as numbers</td>
<td></td>
</tr>
<tr>
<td>- Use fractions to solve problems, including division of fractions</td>
<td></td>
</tr>
</tbody>
</table>
### 2012 Embedded Standards

<table>
<thead>
<tr>
<th></th>
<th>Middle Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>• Understand rational numbers</td>
</tr>
<tr>
<td></td>
<td>• Reason about the size of rational numbers</td>
</tr>
<tr>
<td>7</td>
<td>• Extend understanding of rational numbers</td>
</tr>
<tr>
<td></td>
<td>• Solve problems with rational numbers in any form</td>
</tr>
<tr>
<td>8</td>
<td>• Develop understanding of linear equations</td>
</tr>
<tr>
<td></td>
<td>• Solve and graph linear equations</td>
</tr>
<tr>
<td></td>
<td>• Interpret features of graphed equations</td>
</tr>
</tbody>
</table>
### 2012 Embedded Standards

**Secondary Level**

- Understand quadratic equations
- Solve and graph quadratic equations
- Interpret features of graphed equations
3.NF.1

Stats:
p-value: .765
p-bis: .319

A (12%) and B (9%) were the most popular distractors
Subtract:

\[
\frac{5}{6} - \frac{2}{6}
\]

A) \(\frac{3}{6}\)
B) \(\frac{7}{12}\)
C) \(\frac{7}{6}\)
D) \(\frac{3}{0}\)

4.NF.3a

Stats:
p-value: .697
p-bis: .315

D was the most chosen distractor (27%)
5.NBT.2

Stats:
- p-value: .879
- p-bis: .176

Highest % of distractors was 6% (A)
2012 Embedded Standards
6th Grade Sample Item

The letter Z lies to the right of the letter Y on a number line. The letter Z lies to the right of 0 on a number line. What statement about Y must be true?

A) Y < 0
B) Y > 0
C) Y < Z
D) Y > Z

6.NS.7

Stats:
p-value: .517
p-bis: .233

Distractors about equally chosen (within about 1%)
2012 Embedded Standards
7th Grade Sample Item

The straight lines $\overline{TX}$ and $\overline{UV}$ meet at point V, and form two triangles, $\triangle UTV$ and $\triangle VWX$ shown below.

What is the measure of $\angle VWX$?

A) 43°
B) 71°
C) 92°
D) 109°

7.G.5

Stats:
p-value: .468
p-bis: .321

Distractors about equally chosen (within about 2% of each other)
8.F.2

Stats:
p-value: .302
p-bis: .175

A was the most chosen (35.7%). Others: B (20%), C (14%)
9-12.F-IF.7a

Stats:
p-value: .563
p-bis: .239

A & C about equally chosen
Why do we want items with “good” statistics?

Statistics are not used in isolation. They are all used together to make decisions about items.

- **P-value:** percent of students who answered correctly
- **Point-biserial:** correlation to total score on assessment
- **Foil counts:** how many students picked that answer choice
- **Other measures:** slope, threshold, asymptote

---

P-value: can indicate the difficulty of an item. If an item is too hard or too easy, it may not be at the appropriate grade level.

- Thoughts about “worse than guessing” or < .25? Could mean there is a very attractive distractor (look at foil counts). What makes it attractive? Misconceptions? Poor alignment to teaching practices? Prefer to choose items with p-values > .25 (1/#foils)
- Between .2 and .85 considered “good”

Point-biserial: how well the score on that item relates to the score on the entire test.
- above .3 is “good”, negative values might be a miskey

Foil counts: It is sometimes useful to know what percent of students picked particular distractors. Knowing what wrong answers students picked can help identify misconceptions or common errors (diagnostic foils)
- Less than 10% chosen might be a foil that is unreasonable or not appealing for some reason
Other measures are taken into account, but we won’t go into detail about those right now.
Slope: discrimination (somewhat similar to biserial)
Threshold: difficulty (somewhat similar to p-value)
Asymptote: guessing
2012 Embedded Standards

- Good items, good statistics
  - 5th grade, solving word problems with division
    - Average p: .584
    - Average p-bis: .365

5.NF.3

P range: .243 to .877
P-bis range: .122 to .553

Average p-values of actual 2012 operational test: about .78 across all grades

3rd grade: .88
4th grade: .82
5th grade: .78
6th grade: .80
7th grade: .75
8th grade: .77
HS: .65

KS practice has been to give relatively easy tests.
2012 Embedded Standards

There are 9 pizzas to be divided evenly among 8 families. How much pizza will each family get?

A) $\frac{8}{9}$ pizza
B) 1 pizza
C) $1 \frac{3}{8}$ pizzas
D) $1 \frac{1}{8}$ pizzas

Item Statistics:
P-value: .531
P-bis: .363

5.NF.3

A very close approximation to the “average” stats for this standard.
Lots of various good items mixed in about fractions (and some bad, too).

Example of a “good” item with “bad” stats: Why?
- Although the items were “good”, students may not have been instructed on the items, therefore producing “bad” statistics.

Also, plausible distractors with common misconceptions or student errors.
7.NS.1b

Good? Bad?

Do you use representations like this in your classroom?

What other representations do you use to model addition/subtraction of fractions?
**2012 Embedded Standards**

- **Bad items, bad statistics**
  - 7th Ratios and Proportions
    - Average p: .428
    - Average p-bis: .22

**Bad and Bad: Ratios and Proportions...**

Heavy focus on the part of the standard which emphasizes that the line must intersect the origin.

P range: .059 to .77
Pbs range: -0.01 to .483

What does CETE do with poor questions? Rewrite or tweak. Delete if completely unusable.
7.RP.2a

Most chosen? 64% picked A. Almost equal split between B and C (~15%)

What does CETE do with poor questions? Rewrite or tweak. Delete if completely unusable.
2013 Development

- 635 items targeting CCSSM
- Grades 3 through HS
- Formative use

635 items = about 90 per grade, between 3 and 4 standards at each grade level.

Ultimate destination is the formative engine
# 2013 Embedded Standards

## Elementary Level

<table>
<thead>
<tr>
<th>Grade</th>
<th>Standards</th>
</tr>
</thead>
</table>
| 3     | - Understand attributes of two-dimensional shapes  
       | - Use multiplication to find area  
       | - Classify and compare shapes |
| 4     | - Understand fractions as numbers  
       | - Compare two fractions  
       | - Demonstrate fluency with multiplication  
       | - Classify two-dimensional shapes |
| 5     | - Extend understanding of fractions as numbers  
       | - Use fractions to solve problems, including adding and subtracting fractions  
       | - Understand volume |
## 2013 Embedded Standards

<table>
<thead>
<tr>
<th>Middle Level</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Apply multiplication and division knowledge to ratios&lt;br&gt;• Understand variables in mathematical expressions&lt;br&gt;• Develop an understanding of the coordinate plane&lt;br&gt;• Understand surface area</td>
<td>• Extend understanding of rational numbers&lt;br&gt;• Use rational numbers in any form&lt;br&gt;• Extend 3D geometric knowledge&lt;br&gt;• Compare data distributions</td>
<td>• Understand linear equations&lt;br&gt;• Solve and graph linear equations&lt;br&gt;• Interpret features of graphed equations&lt;br&gt;• Apply knowledge about distances and angles to points and figures on the coordinate plane</td>
<td></td>
</tr>
</tbody>
</table>
# 2013 Embedded Standards

## Secondary Level

- Understand equations and inequalities
- Solve and graph equations and systems
Not clear about “Items reflect the way standards are taught at my school”. Did some choose “neutral” because the CCSS ARE NOT being taught at their school? Not sure.

This question will be revised at the next teacher review in November, and teachers will be asked if they are teaching the standards first, then asked how well the items reflect teaching practices.
EXTRA

6.NS.1

Good? Bad?

40% chose D (add numerators to get 11/3). About equal split between A and B.
The average low temperature in Emporia, KS is \(-5.6^\circ C\) in December and \(-8.3^\circ C\) in January. How much colder is the average temperature in January than the average temperature in December?

A) \(-13.9\)  
B) \(-2.7\)  
C) \(2.7\)  
D) \(13.9\)

**Item Statistics:**
- P-value: .544
- P-bis: .079

**EXTRA**

7.NS.1c

Good? Bad?

P value is OK, but p-bis is not so good. C was chosen by about 32%
Linda works at the market and earns $12.50 an hour. Tony works at the library. The graph below shows the relationship between hours worked and the money that Tony earns.

Which statement correctly compares Linda’s and Tony’s income?

A) Tony earns 6.4 times as much per hour as Linda.
B) Linda earns 0.8 times as much per hour as Tony.
C) Tony earns 1.6 times as much per hour as Linda.
D) Linda earns 1.25 times as much per hour as Tony.

**Item Statistics:**
P-value: .562
P-bis: .403

**EXTRA**

8.EE.5

Good? Bad?

B was the most chosen distractor (23%)
EXTRA

5.NF.7b

Good? Bad?

A next highest choice (24%). C and D about equal.
Thank You!

Alicia Stoltenberg
aliciastoltenberg@ku.edu

Perneet Sood
psood@ku.edu