Michigan Assessment Consortium

Improve learning.
Increase success.

Measurement Issues Inherent in Educator Evaluation

Presentation by the Michigan Assessment Consortium to the OEAA Educator Evaluation Best Practices Conference

April 15, 2011

Why are we here?

- Not just an existential question!
- We have legislation that requires
  - Rigorous, transparent, and fair performance evaluation systems
  - Evaluation based on multiple rating categories
  - Evaluation with student growth, as determined by multiple measures of student learning, including national, state, or local assessments or other objective criteria as a significant factor

Things we're thinking about today

- What is the purpose of the system?
- Some general design considerations for an educator evaluation system
- What non-achievement measures should be part of the system?
- What are some issues involved with non-achievement measures?
- How does the system work for non-teaching staff?

Things we're thinking about today

- How will student achievement be measured?
  - Can't we just use MEAP or MME?
- What types of achievement metrics could/should be used?
- What does research tell us about things that could impact our systems?
- What's a district to do?

What is the purpose of the system?

- Is the purpose simply to identify (and dismiss) low-performing educators?
  - Shouldn't the system really be about promoting universal professional development for educators?
- If the purpose is to promote improvement, how will the system provide feedback to educators?
- What opportunities will be made available for professional growth?

MAC

Designing the system

- How will all of the elements of the system be combined into the overall outcome?
- What will be the nature of the evaluation?
  - Looking at educator “status”
  - Looking at educator “progress” or “growth”
- Who controls the evaluation?
  - Supervisor? Employee? Both?

The Inspection Model

- Supervisor-centered
- Classroom Observation
- Principal/supervisor ratings
- “360 Degree” evaluations
- Parent/student surveys
- Standard achievement test data

The Demonstration Model

- Educator-centered
- Instructional Artifacts
- Teacher self-reports
- Individual achievement test data
- Portfolios

Non-Achievement Measures

- Should they be included?
- If they are, which ones should be used?
- What aspects of “good teaching” or “good leadership” do they capture?
- Do we look at them in a norm-referenced or a standards-based way?

Non-Achievement Measures

- If the non-achievement measures include rating scales:
  - Have the scales been validated?
  - Will raters be trained and monitored?
  - Will Multiple raters be employed?
  - Will inter-rater reliability be established?
  - Will they apply to all educators?
- Will the measures be solely based on the educator or will parent/student perceptions be gathered?

MAC

Non-teaching staff

- Should the same system be used?
- Should the same measures be used?
- Do educators' evaluations impact their supervisors' evaluation?
- What aspects of schooling are non-teaching staff responsible for?

What assessment options do we have in

LOOKING AT STUDENT ACHIEVEMENT?

An underlying assumption

- We have lots of tests that are designed to assess student achievement
- What is not clear is whether these same tests are sensitive to good (or poor) instruction
- Many people take the link from instruction through student achievement to test scores as implicit and obvious
  - These are the things that make measurement specialists nervous!

It has been assumed as obvious that a singular, clear relationship between classroom instruction and test scores exists.

We think that this is dangerous and would ask that we keep this in mind as we move forward.

Measuring Student Achievement

- We have five general options:
  - Rely on MEAP/MME
  - Use other third-party assessments
  - Create educators' own assessments
    - Including portfolios and/or observations
  - Use measures other than tests
  - Some combination of the four above
- Each method has its strengths and weaknesses

Relying on MEAP

Potential Advantages

- Everyone takes it
- Don't have to cut a check for it
- Written to Michigan's curriculum
- Technically strong
- Familiarity – been around for years

Potential Disadvantages

- Not used at every grade
- Not developed for every subject
- May be differentially sensitive to instruction due to sampling
- Not specifically created for teacher evaluation
- "Constructive feedback"?

Third Party Assessments

Potential Advantages

- Choice
- Flexibility
- Technically strong - possibly
- Cost-possibly

Potential Disadvantages

- May be expensive
- May have unknown technical qualities
- May not have been written to your curriculum
- May be differentially sensitive to instruction due to sampling
- Not specifically created for teacher evaluation

District-Created Assessments

Potential Advantages

- Aligned to district offerings
- Can be sensitive to classroom offerings
- May be technically strong
- Familiarity – Created by your staff for your staff

Potential Disadvantages

- Time consuming to develop well
- May be expensive to develop
- May need outside, technical help for development

Non-Test Achievement Measures

Potential Advantages

- May be suitable for all educators
- Avoids the “one size fits all”
- Permits useful data to enter into teacher evaluation

Potential Disadvantages

- Every teacher needs to locate their own measures
- Uneven quality
- Time-consuming to locate and summarize
- Data may not be suitable for educator evaluation

As our legislation requires multiple measures, ideally, we would probably use all four types of measures in our system.

This will add to the complexity and cost of the system, but it will provide the potential to have a more valid system.

The nature of the measures that we choose will be based upon decisions we make as to just what effective teaching is, and what it looks like for us.

Once we have selected our measures....

WHAT SHOULD WE LOOK AT?

What does growth look like?

- Is growth student-centered? (Growth Model)
  - Out 7th grader grew 68 MEAP Scale Points
  - This estimates a trajectory for the student over time and is criterion referenced.
- Is growth teacher-specific? (Value-Added Model)
  - Students in Teacher A's class grow 78 points in a year, whereas students in Teacher B's class grow 61 points.
  - Due to the statistical estimation procedures, this is a norm-referenced viewpoint

Different viewpoints on growth

Briggs, D.C. & Weeks, J.P. (2009). The impact of vertical scaling decisions on growth interpretations. Educational Measurement: Issues and Practice. 28(4). Pp 3-14.

- Individual Student Trajectories (Growth)
  - Computationally fairly simple
  - Summarize/project student achievement over time
  - Requires that the tests used are measuring the same stuff with the same scale
  - Criterion referenced
- Residual Estimation (Value-Added)
  - Computationally more complex
  - Estimate the quantities that support casual inferences about the specific contributions that teachers make to student achievement.
  - Norm-referenced

MAC

What would the differences look like in practice?

- Suppose we used a local test and administered it before and after instruction (pore-post testing)
- We could look at the scores of individual students and see how many had higher post-test scores (Growth Model)
- We could calculate the mean pre-test score and compare that with the mean post-test score (Value-Added Model, simplified)
- Perhaps both are useful

Let’s look at some examples of how achievement data might be used.

WHAT DO WE LIKE/DISLIKE IN EACH?

For the achievement portion of the educator evaluation, a district looks at the percentage of a teacher's current students who were proficient this year and compares it to the percentage of that teacher's students who were proficient the year before.

Positives

- Quick/Cheap

Negatives

- Inappropriate cohort
- Inappropriate for all teachers
- Lots of others…….

MAC

For the achievement portion of the educator evaluation, a fourth grade teacher looks at the number of students who maintained or improved their performance level from last year's fourth grade MEAP to this year's fifth grade MEAP.

Positives

- Quick/Cheap
- Assesses the correct cohort/content
- Uses "the" state assessment

Negatives

- May be instructionally insensitive
- Appropriate for core-content teachers…at least math and reading, at some grades

MAC

For the achievement portion of the educator evaluation, a district looks at the difference in percentile rankings from the last year to this year for students in each teachers' class. An average percentile change is calculated for each teacher and is used to establish growth.

Positives

- Assessment is common across classrooms
- Assessment selected by district…presumption of alignment

Negatives

- May be instructionally insensitive
- May assess different content from one year to the next
- Care must be taken in doing these types of calculations. (NCEs instead of percentiles)

For the achievement portion of the educator evaluation, a teacher gives 4 pre-post tests during the year. For each sequence, the teacher calculates an average change from pre-test to post-test, and looks at the numbers of students whose scores changed in various amounts

Positives

- Chosen so to be instructionally sensitive
- Temporally appropriate
- Easy to understand
- Multiple looks at the data (Growth and VAM)

Negatives

- Potential technical quality issues with tests
- Different teachers in same content/level could choose different tests. Fair?
- Increased reporting and analysis

For the achievement portion of the educator evaluation, a district develops 2 tests to be given pre-post during the year for specific content/grade levels. For each sequence, the district calculates an average change from pre-test to post-test, and looks at the numbers of students whose scores changed in various amount.

Positives

- Built so to be instructionally sensitive
- Temporally appropriate
- Easy to understand
- Multiple looks at the data (Growth and VAM)
- Common across classrooms

Negatives

- Potential technical quality issues with tests if not built thoughtfully/appropriately
- Testing windows and security issues
- Logistics issues for central office

We might like aspects of several of those scenarios to be present in our system.

Thoughtful decisions about *which* tests to use and *how* to use those results will have to be made if the system has any chance of being rigorous, transparent, fair, and valid.

(The system must) “take into account data on student growth as a significant factor.”

WHAT DOES SIGNIFICANT MEAN?

Significant?

- Supt. Flanagan has said he thinks 40-60% constitutes significant.
- Would you feel that a 10% cut to your pay is significant?
- Classically trained statisticians hear significant and automatically thing 5%
  - (.05, $\alpha < .01$ ☺)
- Perhaps we shouldn't decide how much is "significant" until we know what else is in the system.

Significant, revisited…

- Perhaps out “growth as a significant factor” should be answered in the context of the other elements chosen to be in the system.
- If we have confidence in the quality of the non-achievement instruments, “growth measures” may have a lower weight in the scoring system.
- On the other hand, if we think our achievement measures are better than our non-academic measures, we might want growth to count more.

MAC

As if that weren't enough...

WHAT ELSE SHOULD WE BE THINKING ABOUT?

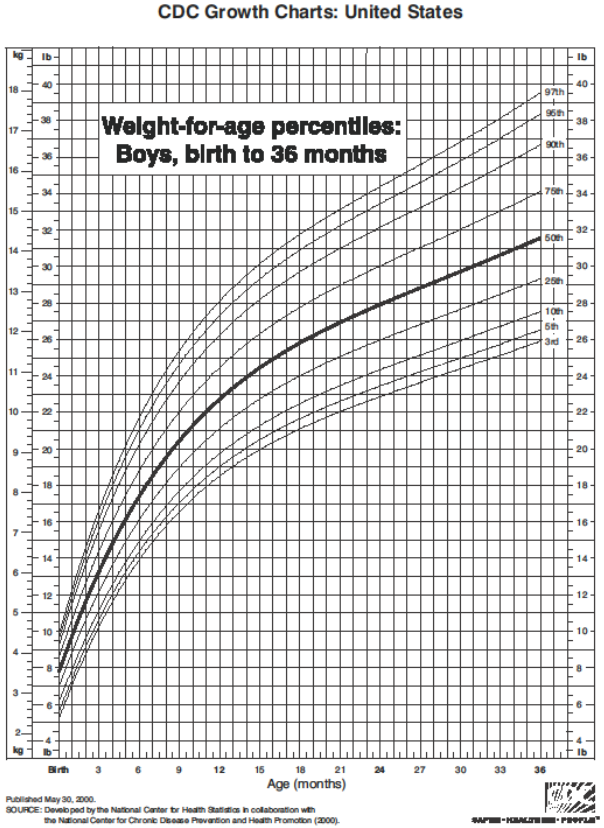
Additional things to consider…

- Should teachers be able to self-select measures? Is that fair? How should they be weighted?
- How will principals, counselors, district administrators, librarians, …etc. be evaluated?
  - Hierarchical Linear Models? (!) Transparent?
- What time frames are appropriate?
  - Multi-year, action research projects possible?

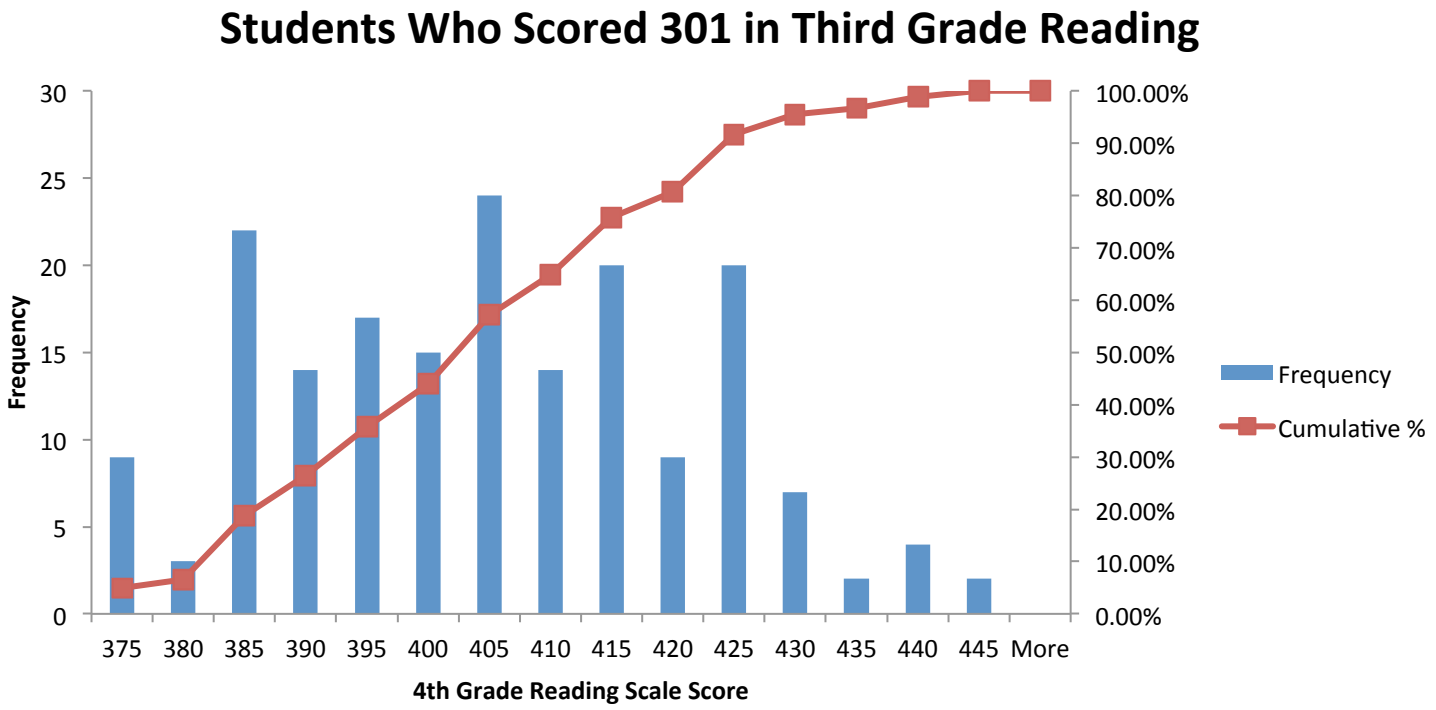
If we have time...

GROWTH REVISITED

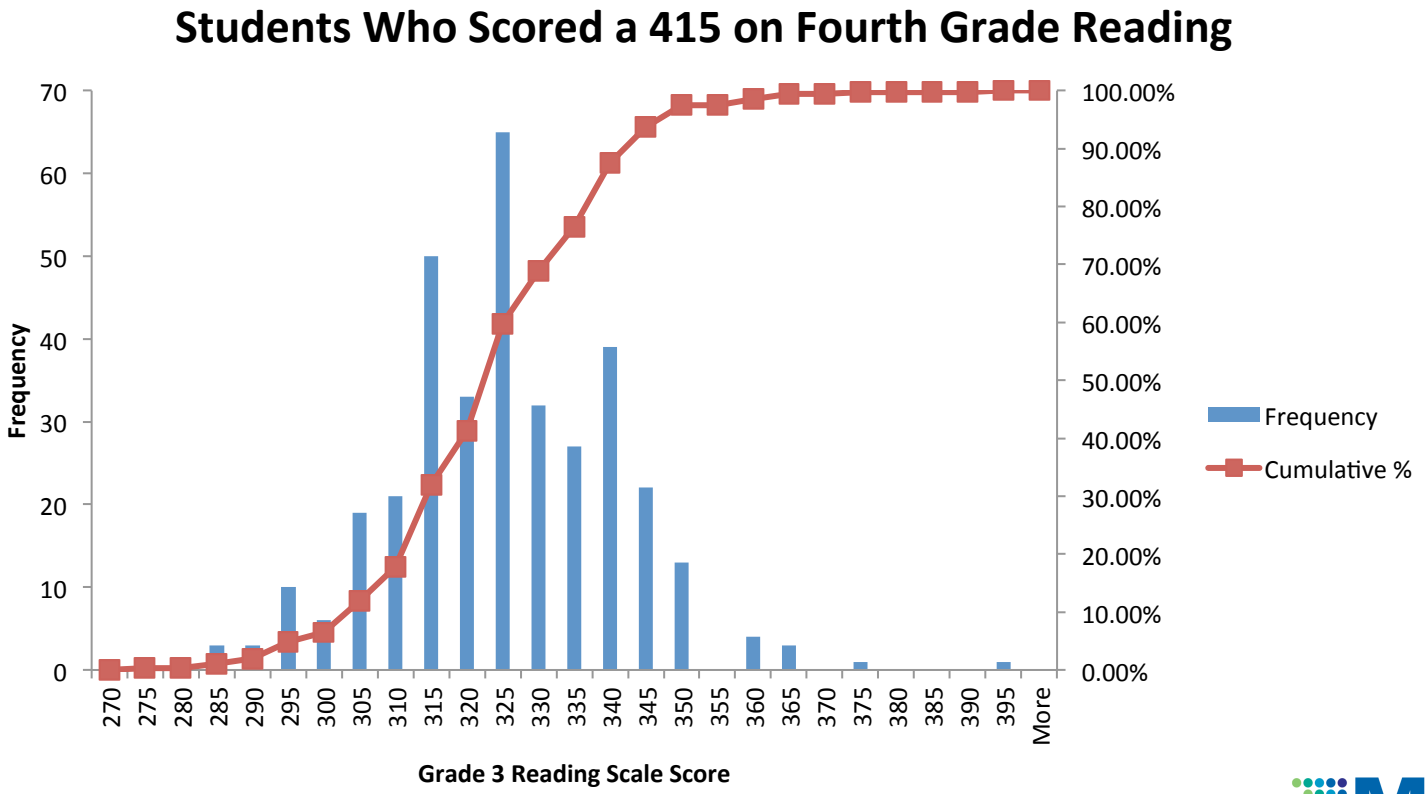
If you're a parent, you probably recognize this....



MEAP Growth Charts (Reading)



MEAP Growth Charts (Reading)



MEAP Growth Charts (Reading)

31st Percentile Growth

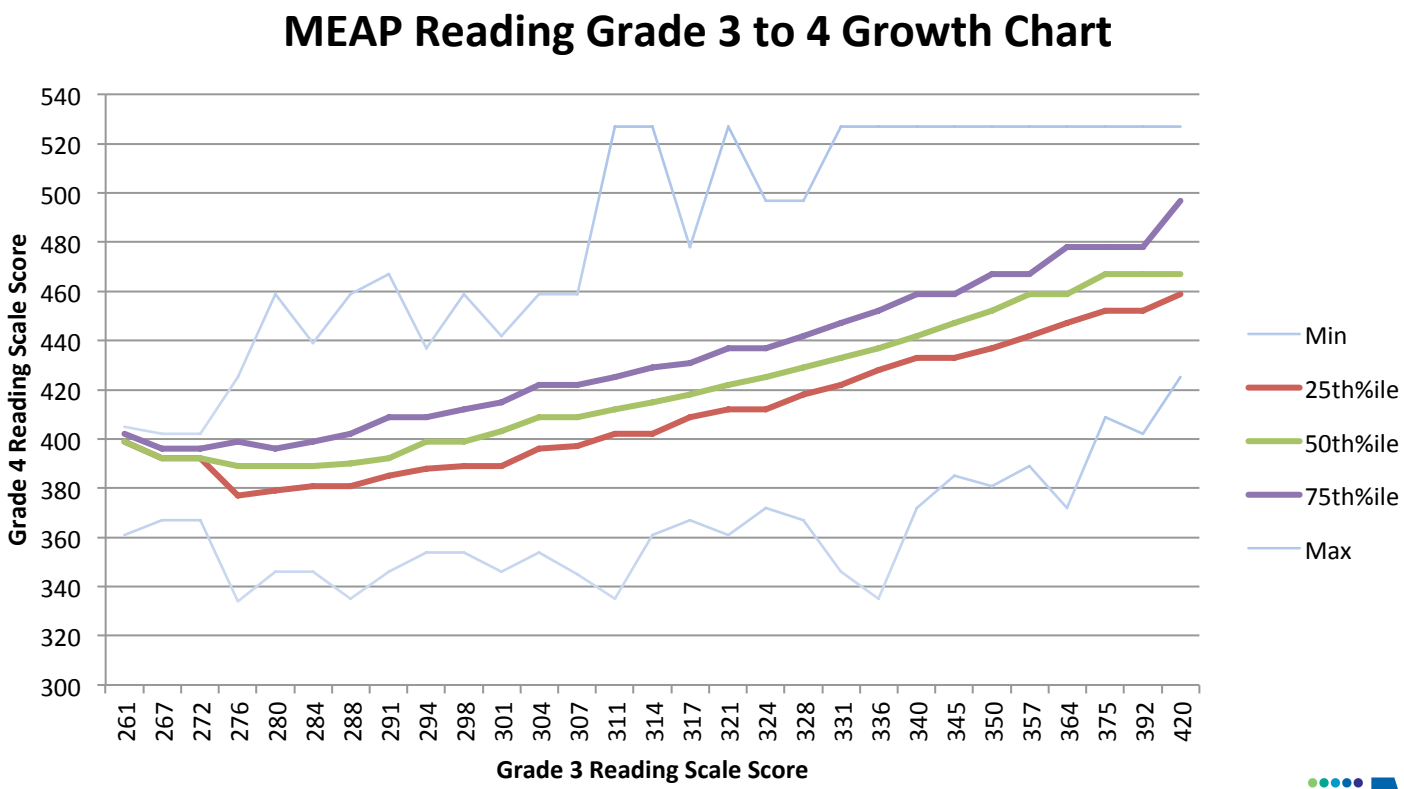
G3 Reading	G4 Reading	Difference
272	377	105
298	392	94
324	418	94
328	422	94
340	433	93
364	452	88

70th Percentile Growth

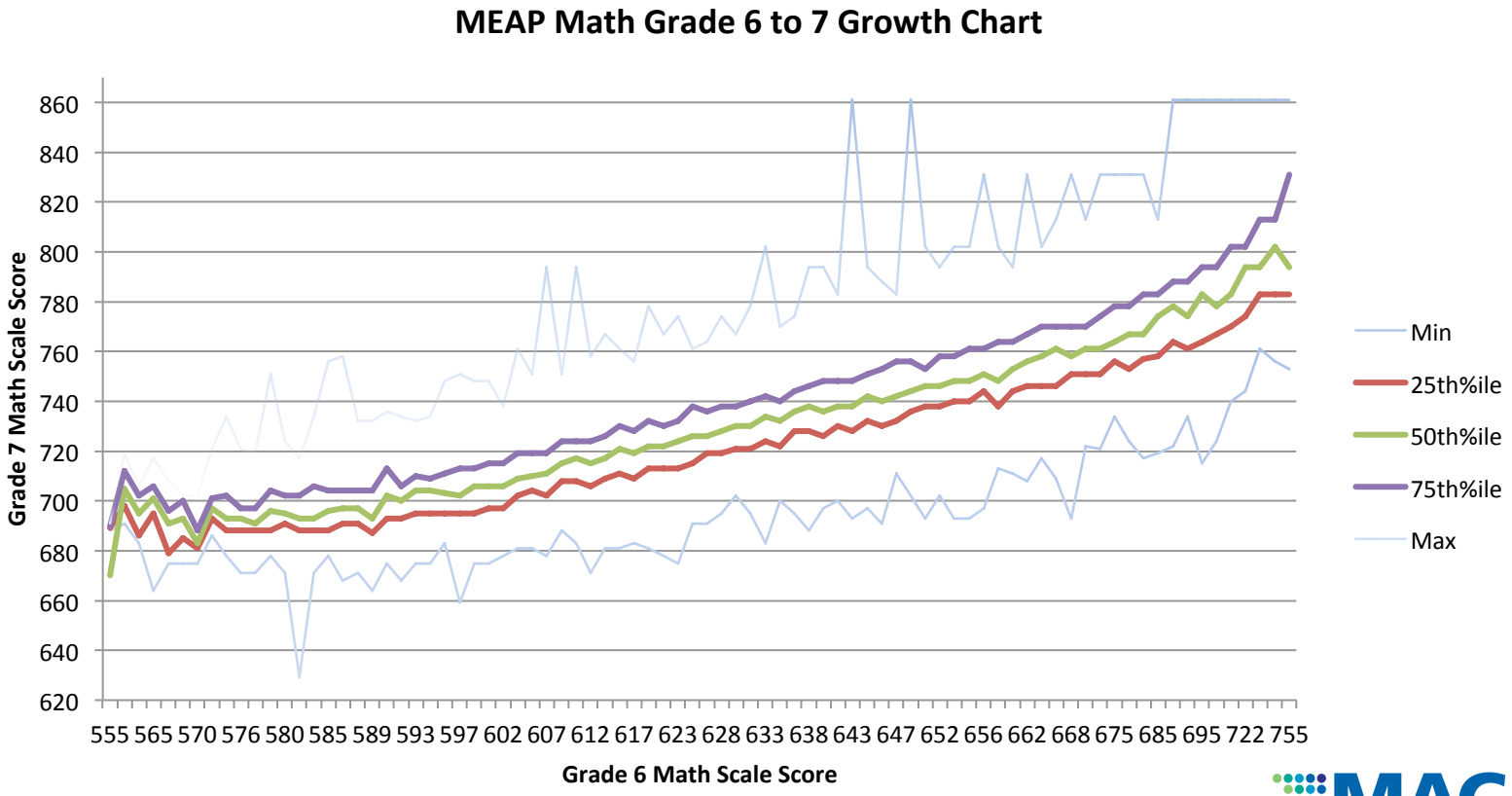
G3 Reading	G4 Reading	Difference
267	396	129
276	396	120
298	409	111
301	412	111
304	415	111
317	429	112

MAC

MEAP Growth Charts (Reading)



MEAP Growth Charts (Math)



Student Growth Percentiles

Betebenner, D. (2009). Norm- and criterion reference student growth. Educational Measurement: Issues and Practice. 28(4). Pp 42-51.

Advantages

- Based on “reality”
- Conceptually familiar
- Growth is independent of status
- Some 20 states are using student growth percentiles in some form
- Can be used to project growth

Disadvantages

- Requires LARGE data sets
- Complex mathematics
  - Sparse N techniques
  - Quantile Regression
  - Transparent?
- How does it fit in with a “criterion referenced” system?

MAC

Many Thanks!

Jim Gullen
- james.gullen@Oakland.k12.mi.us

Ed Roeber
- roeber@msu.edu