How Certain Can You Be?

A presentation for teachers and administrators with questions about the statistics of using student growth for teacher evaluation.

Hello

- I teach high school mathematics in southwest, tribal Alaska.
- I worked for two years in a lab that used mathematics to study how humans make decisions under conditions of uncertainty.
- Our district was among the first to use NWEA MAPS data for a student-growth component of teacher evaluation; we made a noble effort and also some mistakes.

Hello

A quote I like about uncertainty:

“I returned, and saw under the sun, that the race is not to the swift, nor the battle to the strong, neither yet bread to the wise, nor yet riches to men of understanding, nor yet favour to men of skill; but time and chance happeneth to them all.”

--Ecclesiastes 9:11

Our Dilemma

Teacher 1: A teacher is testing their five students and receives the following results:

Student	Score
1	-.31
2	-.15
3	+.31
4	-.93
5	-.15

- Results: 80% failed to meet growth projections.

Our Dilemma

The teacher comes to you and says,

“Sam and Nora (two students who showed insufficient growth) have grown so much in their classwork. But Sam came to me after the test and said that she hadn't eaten breakfast that morning, and Nora was having trouble staying awake all day.

“I don't think the test results really reflected these students' abilities.”

Our Dilemma

Teacher 2 is testing thirty students.

- Six of their students make their growth projections and 24 fail to make their growth projections.
- Results: 80% failed to meet growth projections.

Our Dilemma

The teacher comes to you and says,

“Most of my students have shown so much growth in their school work. Isn't it possible that a dozen of them came in hungry that day, or tired? Or maybe by bad luck they got a few extra questions wrong?

“I don't think the test results really reflected my students' abilities.”

Our Dilemma

After Seeing:

5 Students
80% “failed”

30 Students
80% “failed”

Are You Now Confident Enough To:

- Withhold incentive pay?
- Tell the state the teacher is “bad”?
- Remove the teacher from the classroom?

Less Certain

- Withhold incentive pay?
- Tell the state the teacher is “bad”?
- Remove the teacher from the classroom?

More Certain

Our Dilemma

After Seeing:

5 Students

80% “failed”

30 Students

80% “failed”

Are You Now Confident Enough To:

- Withhold incentive pay?
- Tell the state the teacher is “bad”?
- Remove the teacher from the classroom?

Less Certain

- Withhold incentive pay?
- Tell the state the teacher is “bad”?
- Remove the teacher from the classroom?

More Certain

Our Dilemma

Our Dilemma

Punchlines:

- Student growth measures present an opportunity to bring quantitative accountability to school administration.
- And yet, too simple of a method for reporting aggregate student growth can neglect this feeling of uncertainty we've identified.
- Failure to navigate this dilemma may harm teacher morale and cost us opportunity and money.

Outline

- Hello
- Our Dilemma
- Measurement
- Measurement Error & Uncertainty
- A Year's Growth
- The Standard Growth Score
- Statistical Models for Teacher Evaluation
- Dealing with Uncertainty
- Goodbye

Measurement

- We can measure the height of a teacher...

  ...by standing them against a door-jam, putting a mark at the top of their head, and measuring the height of that with a measuring tape.

Measurement

- We can measure the volume of a teacher...

  ...by filling a barrel to the brim with water, forcing the teacher to climb inside, and measuring the amount of water displaced.

Measurement

- We can measure the teaching ability of a teacher...

  ...by measuring the academic ability of a group of students, placing those students in a room with the teacher for a school-year, and measuring the group again post-anum: we attribute any difference in the two measurements to the teacher's abilities.

Measurement

- We can measure the teaching ability of a teacher...

  ...by measuring the academic ability of a group of students, placing those students in a room with the teacher for a school-year, and measuring the group again post-anum: we attribute any difference in the two measurements to the teacher's abilities.

Measurement

First we must assume:

- That the academic ability measured by MAPS aligns with your institution's instructional goals.
- That the academic ability measured by the MAPS subject area test aligns with the instructional goals of the teacher to whom you attribute that test's results.
- That growth as measured by MAPS over one academic year should be attributed to the instruction received during that academic year.

Measurement

First we must assume:

- That the academic ability measured by MAPS aligns with your institution's instructional goals.
- That the academic ability measured by the MAPS aligns with your testing window with the instructional goals of the teacher to whom you attribute those test's results.
- That growth as measured by MAPS over one academic year can be attributed to the instruction received during that academic year.

We do not address these assumptions in this presentation, but the validity of your teacher evaluation hinges on them.

Measurement

Punchlines:

- We can try to measure any attribute of a teacher.
- The way we've decided to measure teaching ability forces us to measure student growth first.
- So let's discuss measuring student growth.

- But don't forget: you must consider key assumptions that will hinge the validity of these measurements.

Outline

- Hello
- Our Dilemma
- Measurement
- Measurement Error & Uncertainty
- A Year's Growth
- The Standard Growth Score
- Statistical Models for Teacher Evaluation
- Dealing with Uncertainty
- Goodbye

Measurement Error & Uncertainty

- I am 5' 11.5”, to the nearest half inch.
- Imagine an engineer measuring my height with a laser and declaring that I am 5' 11.63”, to the nearest hundredth of an inch.
- What would you say about my measuring ability?

Measurement Error & Uncertainty

- For simple measurements like height, or weight, or volume, we talk about measurement error by saying, “to the nearest...”

- For complicated measurements like academic ability and teaching ability, measurement error is more complicated.

Measurement Error & Uncertainty

An example from the field of sports:

Karen wants to join her track and field team. The coach says that she can try out for javelin throw as follows:

- Every morning for a week she will throw the javelin one time.
- If her *average* distance is long enough, she will be allowed on the team.

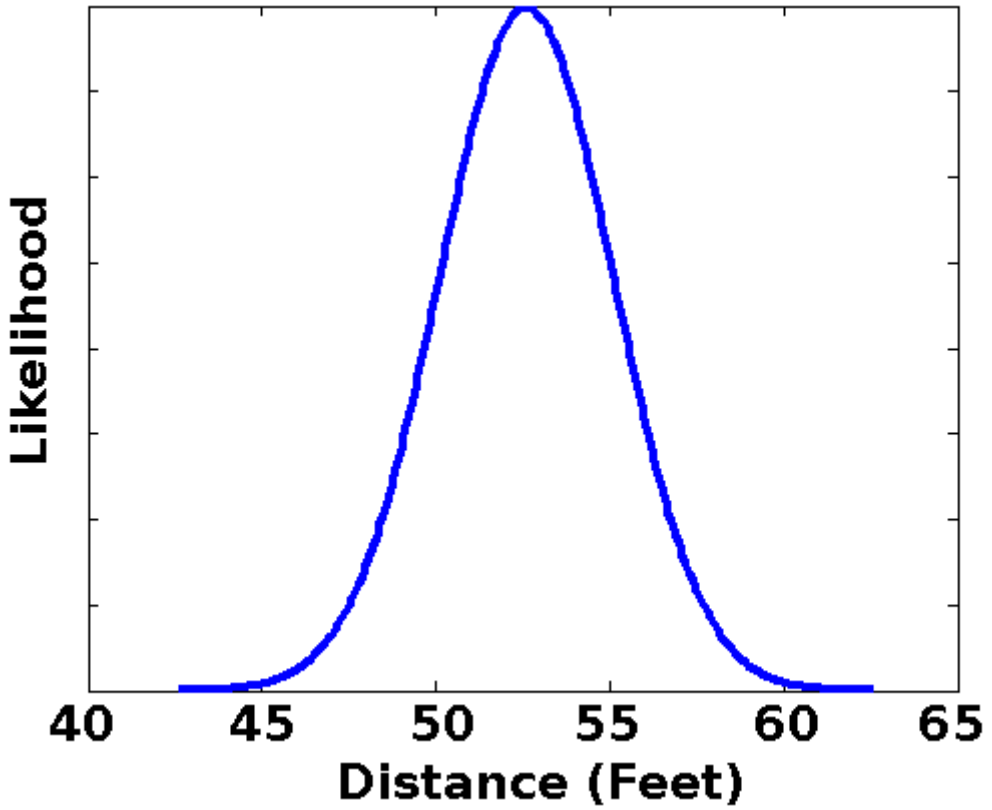
Measurement Error & Uncertainty

Karen's throws:

Day	Distance	Average So Far
1	52ft	52ft
2	61ft	56.5ft
3	48ft	53.6ft
4	43ft	51ft
5	60ft	52.8ft
6	48ft	52ft
7	56ft	52.6ft

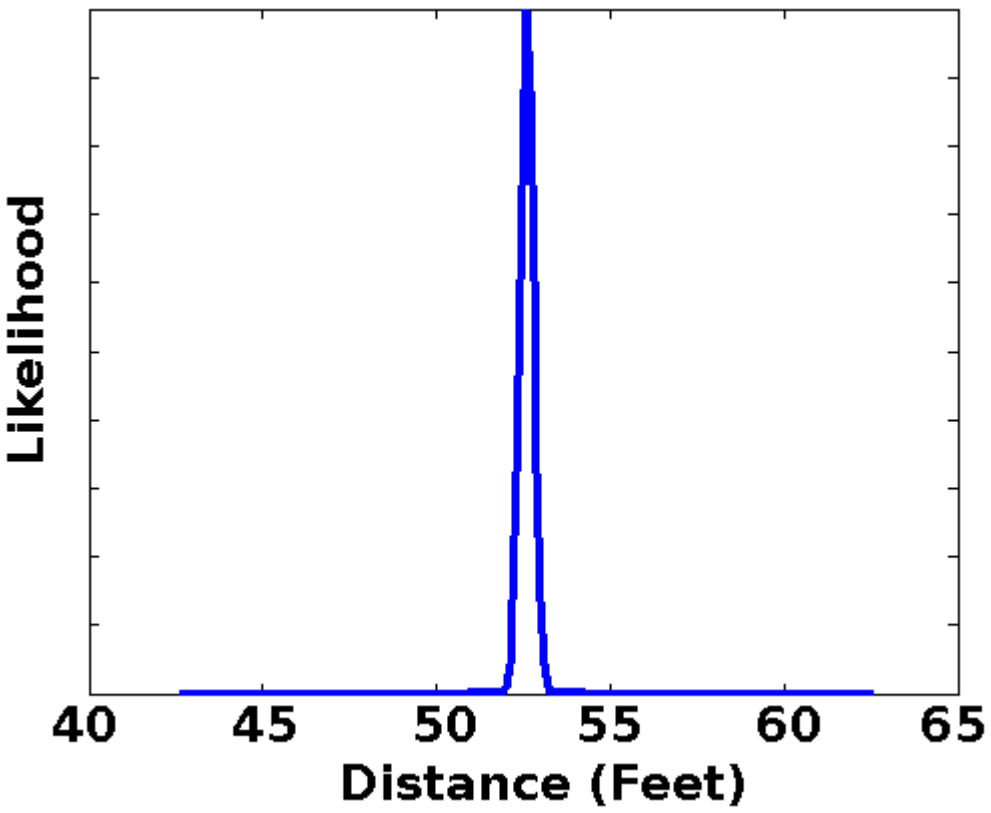
Measurement Error & Uncertainty

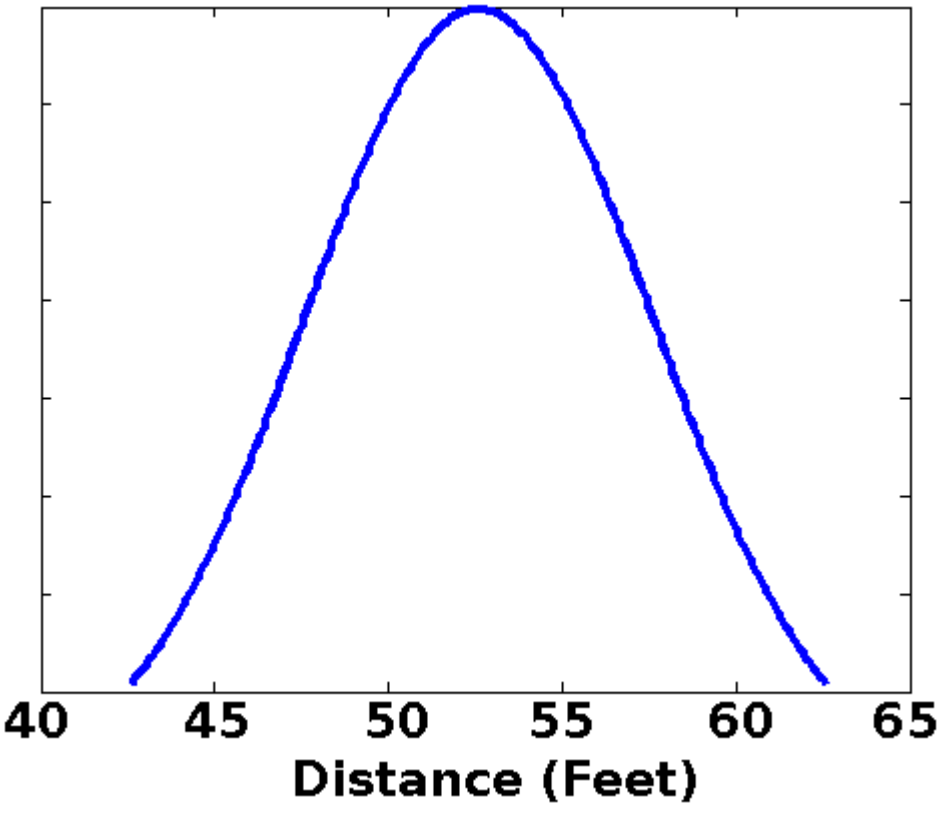
A representation of likely values of Karen's ability.



Measurement Error & Uncertainty

Similar figures for other athletes





Measurement Error & Uncertainty

Sources of measurement error for MAPS:

- Poor sleep.
- Missing breakfast.
- Girl crazy.
- Boy crazy.
- Bad luck.
- Harsh words from another teacher.
- Change in eyesight.
- Academic motivation.

-or-

- Good luck on last year's test.

Measurement Error and Uncertainty

Punchlines:

- Any academic test will include measurement error.
- Because of measurement error, we will have inescapable uncertainty about students' true academic ability.

Preview:

- This means we will have inescapable uncertainty about teacher performance (but we will learn how to quantify that uncertainty).

Outline

- Hello
- Our Dilemma
- Measurement
- Measurement Error & Uncertainty
- A Year's Growth
- The Standard Growth Score
- Statistical Models for Teacher Evaluation
- Dealing with Uncertainty
- Goodbye

Roland Napoka

- Starting Second Grade
- 180 RIT Score in mathematics.
- Super cute.

A Year's Growth

"All growth projection *Targets* and references to *Typical Growth* shown on reports are the *Average Growth* exhibited by the students in the same grade who started at the same RIT level in the sample set of the NWEA RIT Scale Norms study."

– NWEA, http://www.nwea.org/node/4355

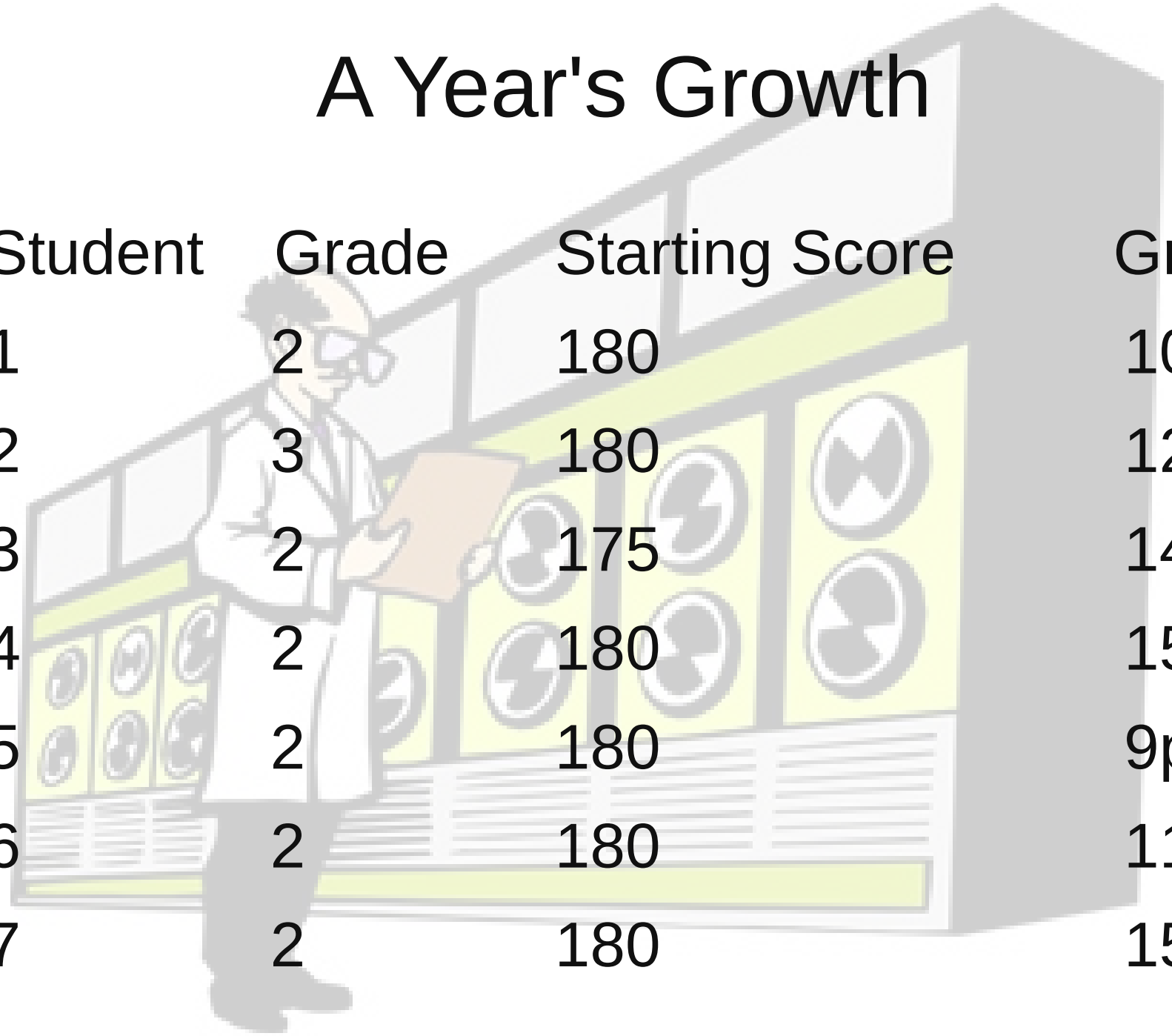
A Year's Growth

"...*Average Growth* exhibited by the students in the same grade who started at the same RIT level..."

- **Historical Cohort**

Let's ask the computer to find Roland's historical cohort.

A Year's Growth



Student	Grade	Starting Score	Growth
1	2	180	10pts
2	3	180	12pts
3	2	175	14pts
4	2	180	15pts
5	2	180	9pts
6	2	180	11pts
7	2	180	15pts

A Year's Growth

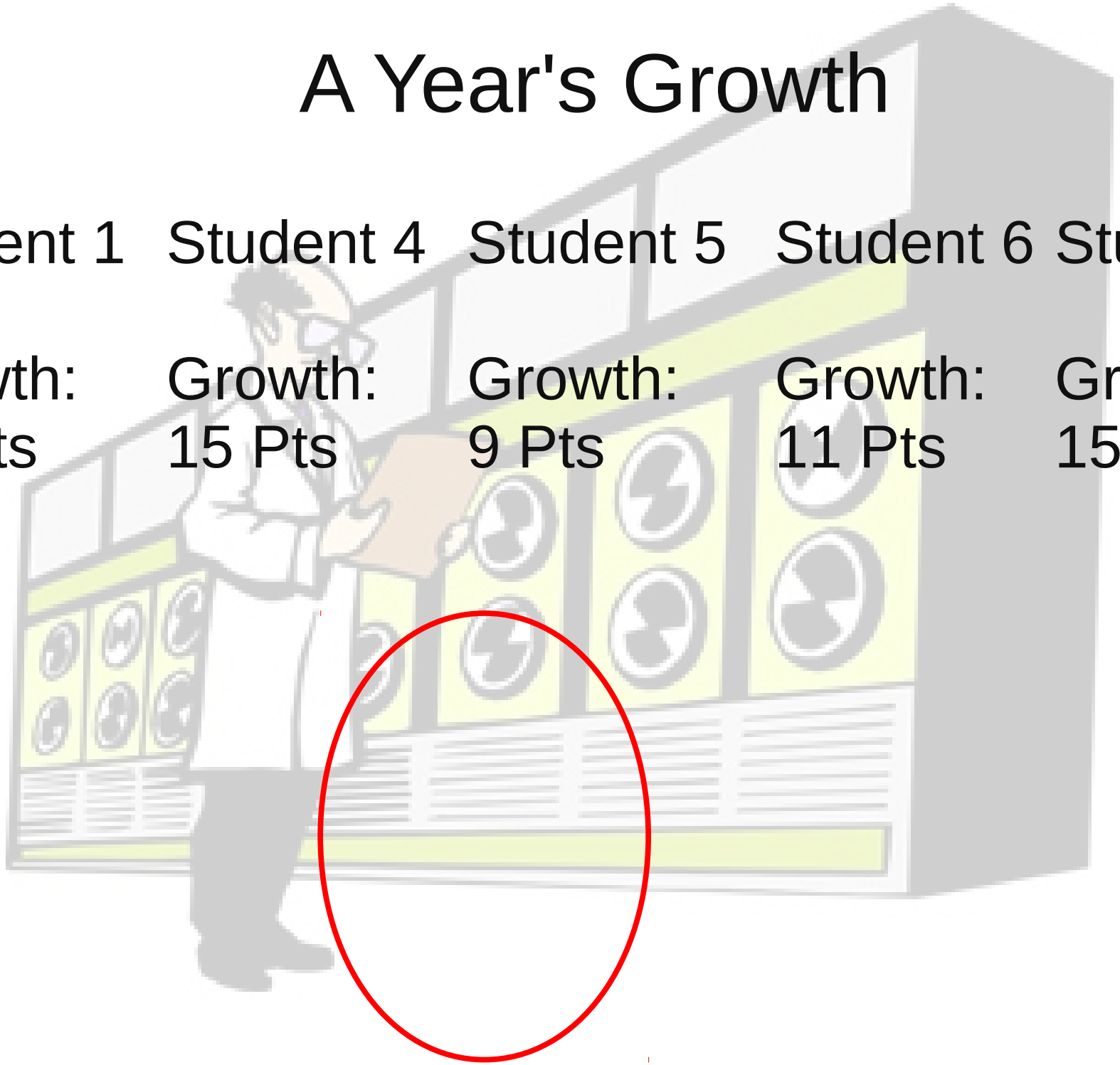
"All growth projection *Targets* and references to *Typical Growth* shown on reports are the *Average Growth* exhibited by [the student's historical cohort]."

- **A Typical Year's Growth**

Let's ask the computer to find a "typical year's growth" for Roland.

A Year's Growth

Student 1	Student 4	Student 5	Student 6	Student 7
Growth: 10 Pts	Growth: 15 Pts	Growth: 9 Pts	Growth: 11 Pts	Growth: 15 Pts



A Year's Growth

- A typical year's growth *is* the AVERAGE growth demonstrated by the student's historical cohort over the course of a year.
- AVERAGE

A Year's Growth

- A typical year's growth *is not* a standards-referenced bar that every student should be expected to meet.

A Year's Growth

Punchlines:

- If you didn't have to deal with chance and uncertainty, you would expect 50% of your students to exhibit above average growth and 50% of your students to exhibit below average growth.

- All references in NWEA to targets and typical growth are growth averages calibrated to your students. Under average conditions, half of your students will meet these targets and half will not.

Outline

- Hello
- Our Dilemma
- Measurement
- Measurement Error & Uncertainty
- A Year's Growth
- The Standard Growth Score
- Statistical Models for Teacher Evaluation
- Dealing with Uncertainty
- Goodbye

The Standard Growth Score

Consider these two students:

Jamie:

- Most students in his historical cohort grow 5-20 points.
- On average, these students grow 12 points.
- Jamie grows 19 points.
- Good job Jamie.

Julie:

- Most students in her historical cohort grow 11-13 points.
- On average, these students grow 12 points.
- Julie grows 19 points.
- Wow.

The Standard Growth Score

Punchlines:

- Captures the “exceptionalness” of the student score.
- Similar to a percentile score, but easier to work with mathematically.
- Captures the most information about the student's achievement, relative to their historical cohort, except a measure of uncertainty.

The Standard Growth Score

From the NWEA ASG Calculator:

Status Norms		
Predicted Mean of Grade Level RIT Score	220	225
Predicted SD of Grade Level RIT Score	15.41	16.08
Student's Status Percentile	10	6

Growth Norms	
Observed Student Growth	0
SEM of Observed Student Growth	4.67
Growth Projection	6
SD of Growth	6.41
Conditional Growth Index	-0.91
Student's Growth Percentile	18

Negative numbers are bad, positive numbers are good, zero is average growth.

Outline

- Hello
- Our Dilemma
- Measurement
- Measurement Error & Uncertainty
- A Year's Growth
- The Standard Growth Score
- Statistical Models for Teacher Evaluation
- Dealing with Uncertainty
- Goodbye

Statistical Models for Teacher Evaluation

We have to decide what it means to be a “good” teacher.

We consider two options:

- The more students that demonstrate a typical year's growth, the better the teacher.
- The higher the average “exceptionalness” (standard growth score) of students, the better the teacher.

Statistical Models for Teacher Evaluation

Model A: The more students that demonstrate a typical year's growth, the better the teacher.

Advantage:

- Promotes equity in outcome.

Disadvantage:

- Ignores some available aspects of teacher performance.

Statistical Models for Teacher Evaluation

Model B: The higher the average standard growth score of students, the better the teacher.

Advantage:

- Makes use of all the most comprehensive information.

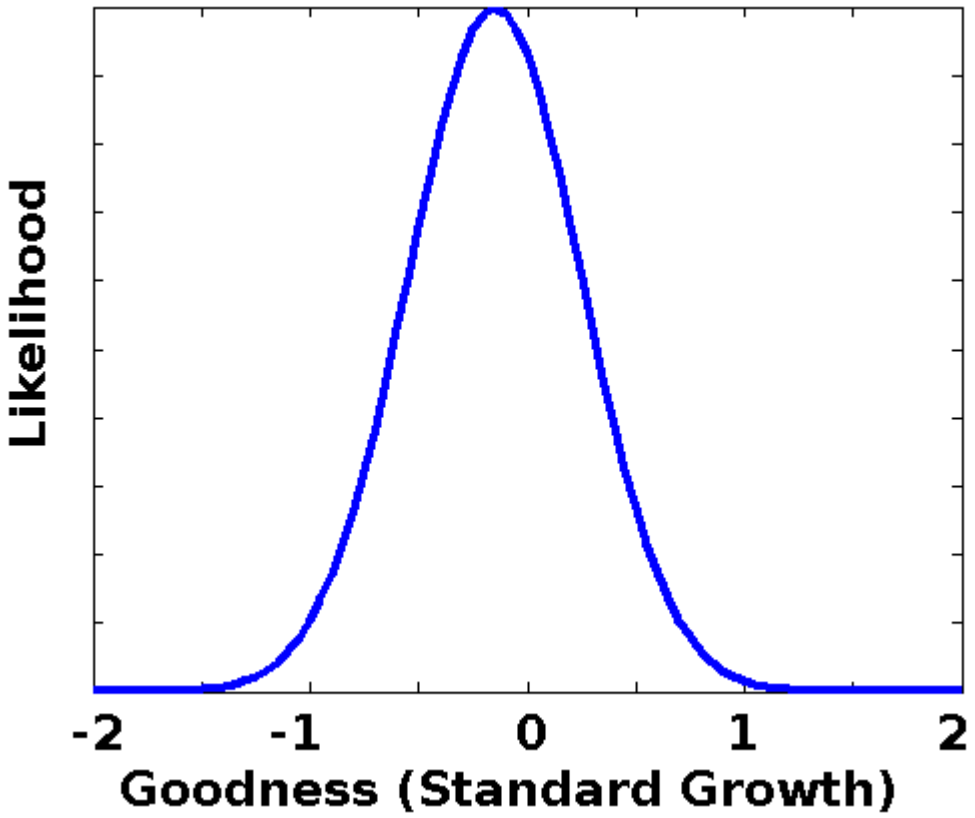
Disadvantage:

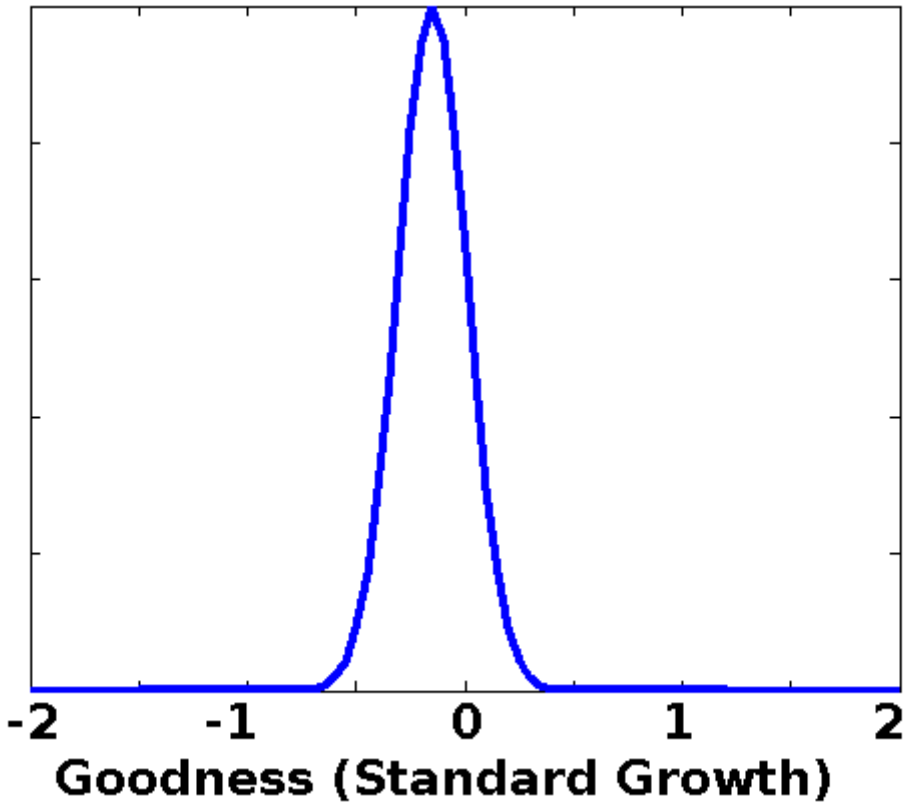
- Could encourage teachers to accept unequal classroom performance.

Statistical Models for Teacher Evaluation

Teacher 1

Teacher 2





Statistical Models for Teacher Evaluation

Crib notes for implementing these models.

- Model A is a simple Bayesian test incorporating the SEM of observed student growth, as taken from the ASG calculator.
- For Model B, we calculate the median student standard growth score because student growth scores are ordinal.

Statistical Models for Teacher Evaluation

Crib notes (continued).

- We use propagation of uncertainty to estimate the standard error for calculating the mean standard growth score, and then multiply by 1.25 to estimate the standard error for calculating the median.
- These models are back-of-the-envelope, developed to demonstrate the potential magnitude of statistical uncertainty.

Statistical Models for Teacher Evaluation

Punchlines:

- You have to decide, from multiple valid options, what constitutes a “good” teacher.
- Neither of our options give us enough quantitative certainty to say that either teacher is definitely bad.

Outline

- Hello
- Our Dilemma
- Measurement
- Measurement Error & Uncertainty
- A Year's Growth
- The Standard Growth Score
- Statistical Models for Teacher Evaluation
- Dealing with Uncertainty
- Goodbye

Dealing with Uncertainty

Bad quantitative result triggers a qualitative review.

Advantage:

- Gives decision-making authority to seasoned educators.

Disadvantage:

- May not comply with state requirements.

Dealing with Uncertainty

When facing uncertainty, give teachers the benefit of the doubt.

Advantage:

- Gives states/authorities a “one number” student growth component.

Disadvantages:

- Does not eliminate uncertainty.
- Could unfairly advantage teachers with fewer students.

Dealing with Uncertainty

Integrate more tests.

Advantage:

- More “statistical power”.

Disadvantages:

- Difficult to find tests as appropriate to measuring student growth as MAPS.
- Drain on student time.

Dealing with Uncertainty

Evaluate schools and districts, not teachers, based on test results.

Advantage:

- More “statistical power”.

Disadvantage:

- Might not satisfy state requirements.

Outline

- Hello
- Our Dilemma
- Measurement
- Measurement Error & Uncertainty
- A Year's Growth
- The Standard Growth Score
- Statistical Models for Teacher Evaluation
- Dealing with Uncertainty
- Goodbye

Goodbye

http://www.jaschilz.net/essays.php