AMERICAN INSTITUTES FOR RESEARCH®

Florida's Race to the Top

Student Growth Implementation Committee (SGIC)

University of Central Florida

Orlando, FL

May 19-20, 2011

Meeting Agenda

Thursday, May 19, 2011

- 8:00 am- Coffee
- 8:30 am- Welcome, Agenda Overview, Process of Selecting a Model
- 9:30 am- Results of the Value-Added Model Evaluation
- 10:30 am- Break
- Noon- Lunch on your own
- 1:15 pm- Results of the Value-Added Model Evaluation
- 5:00 pm- Adjourn

Meeting Agenda

Friday, May 20, 2011

- 7:30 am- Coffee and Informal Conversation with AIR Team
- 8:30 am- Review of Thursday's Discussion
- 9:30 am- Discussion on Model Selection, Business Rules and Variables
- Noon- Lunch on your own
- 1:15 pm- Select a model to recommend to the Commissioner
- 3:15 pm- Break
- 3:30 pm- Next Steps
  - Webinar, May 25, 2011 from 4:30-6:30 pm
  - Initial White Paper Outline
  - Course Code Directory Discussion
- 5:00 pm- Adjourn

AIR

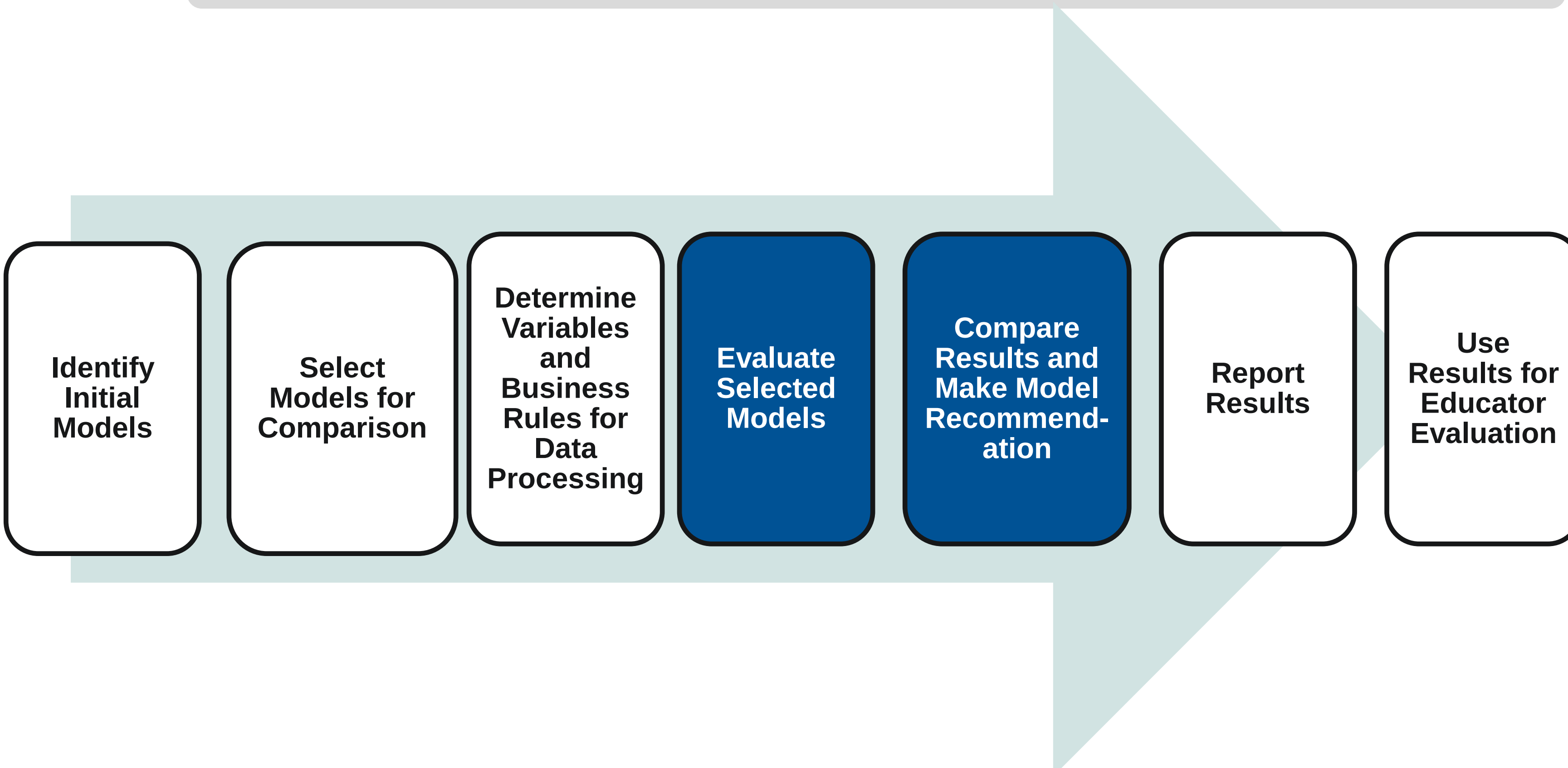
SGIC Purpose and Expectations

The purpose of the SGIC is to provide input, seek feedback, and present recommendations to the state in the development and implementation of teacher-level student growth models.

The SGIC is not responsible for final decisions regarding the adoption of a state model or the district models.

The process for providing input, feedback, and recommendations to the state will continue over the four years of the project.

Focus Steps for May 19-20 Meeting



Meeting Goal

Select a value added model and variables to recommend to the Commissioner for use in teacher evaluation by June 1, 2011.

Selecting a model

- AIR does not advocate for or against any particular model.
- AIR's role is to facilitate Florida's conversation and choice of model by:
  - **Identifying** different VAM models for SGIC to consider
  - **Comparing** the selected model results against a set of empirical and policy criteria
  - **Reporting** these findings to the state, the SGIC, and other advisory groups for consideration

AIR

Background

- At the April 4-5, 2011 Student Growth Implementation Committee (SGIC) meeting at the University of Central Florida, the SGIC selected three value-added models for AIR to evaluate.
  - Allow student characteristics and prior achievement scores
    - fixed and random effects
  - Include only prior achievement scores
  - Sustained differences model
- Three variables, Students with Disabilities, English language learner status, and attendance, will be evaluated in these models as determined and defined by the SGIC.
- The SGIC proposed several additional variables for consideration in the evaluation of the models: gifted, class size, age, mobility, homogeneity, school effect.

Value Added Evaluation Milestones

- Eight different value-added models were presented to the SGIC for discussion
- The SGIC selected three models for AIR to evaluate and provided AIR discretion to develop variants of these models
- SGIC provided guidance and direction on business rules
- SGIC selected several variables to be evaluated, SWD, ELL, gifted, attendance, age, mobility, homogeneity, and class size
- AIR evaluated the models and variants and is presenting the results of the models at the May 19-20 meeting
- SGIC must make a recommendation to the Commissioner by June 1

Value added models evaluated

The difference model expects the same amount of growth from each student in an achievement level

Differences Model: Expects students who score the same to continue to score the same and assumes the same amount of growth for each student in each achievement level.

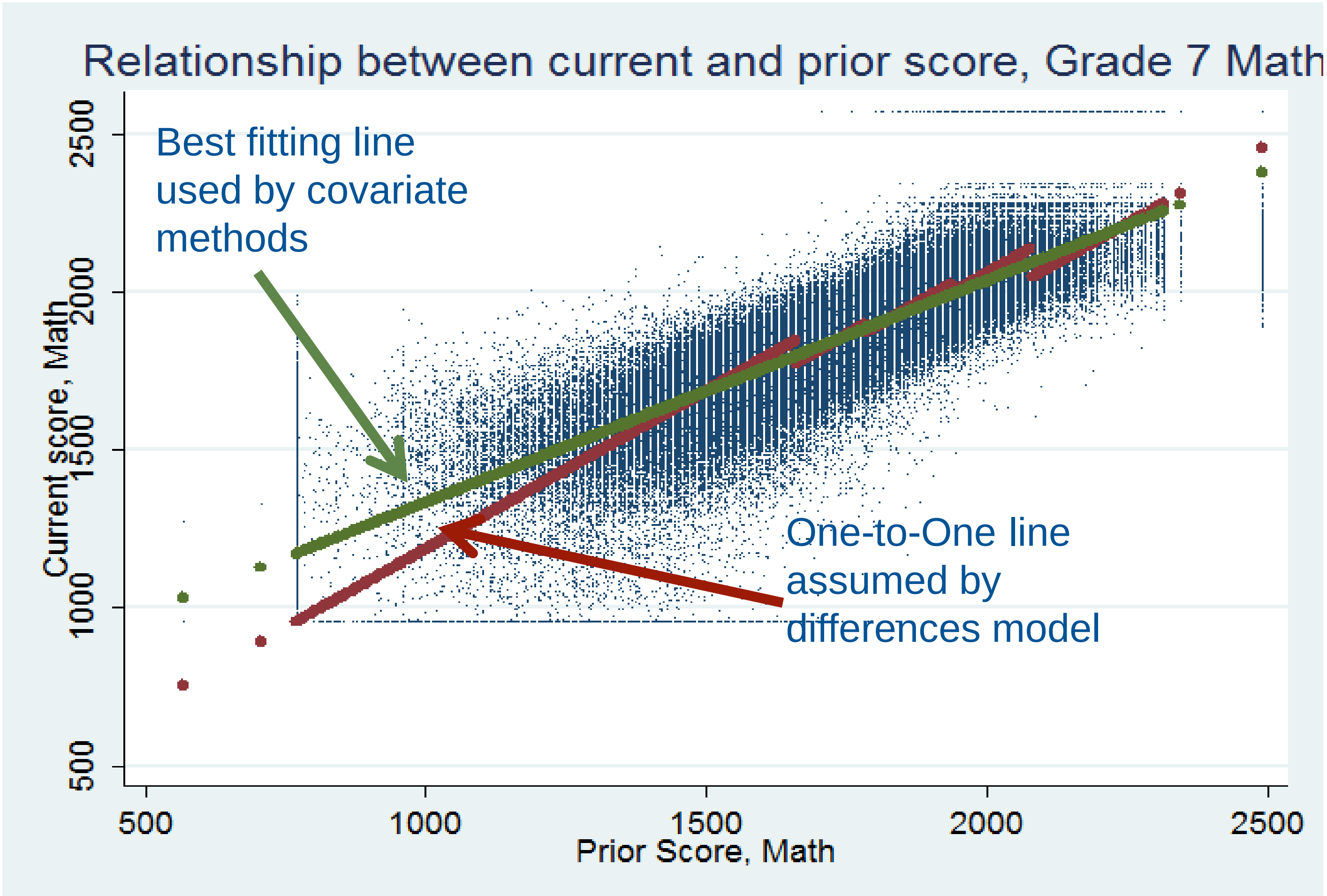
The covariate model allows expected growth to vary within achievement level

Covariate adjustment models: These models expect students who score the same in prior years to score the same the next year. Expected growth may vary within achievement level.

How does the differences model differ from the covariate models?

The differences model predicts more growth in the top end of each achievement level than is typically seen

It also predicts less growth at the bottom of achievement level 1



Value added models evaluated

1) A two-level model that includes only teacher effects, control for one year prior achievement, and control variables for ELL, SWD, and attendance and is estimated with random effects
   a) Model 1 with two years of prior achievement
2) Model 1 estimated with fixed effects
3) A core three-level model that that includes teacher and school effects, control for two years prior achievement, and varies as to which variables are included.
   a) No additional variables
      i. Model 3a with only one year of prior achievement
   b) ELL, SWD, and attendance. Use one or two years of prior achievement depending on the whether the earlier year matters in a 3a. *Note: Use two years prior achievement*
   c) ELL, SWD, attendance, class size, homogeneity of class composition, mobility, difference from modal age. Use one or two years of prior achievement depending on the whether the earlier year matters in a 3a. *Note: Use two years prior achievement*
4) Differences model (described in a separate specifications document)

Summary of models

Model Name	Effects estimated	Years of Student Data	Other variables
Model 1	Teacher only, Random	1	SWD, ELL, and attendance
Model 1a	Teacher only, Random	2	SWD, ELL, and attendance
Model 2	Teacher only, Fixed	1	SWD, ELL, and attendance
Model 3a	Teacher and school	2	None
Model 3ai	Teacher and school	1	None
Model 3b	Teacher and school	2	SWD, ELL, and attendance
Model 3c	Teacher and school	2	SWD, ELL, attendance, gifted, class size, homogeneity, mobility, age
Model 4 Differences	Teacher	1	Achievement Level

Fixed or random effects

- These are minor statistical nuances with some possible impact on the data.
- These are two different ways of estimating the same thing.
- Fixed and random effects are known to converge to the same value as the number of students in a class gets larger.
- Rationale for testing assumption: To see whether teacher effects are similar between different estimation approaches.

Addition of student characteristics

- Recall that a VAM is designed to mitigate the fact that there is an unequal distribution of student proficiency and characteristics across classes.
- There is some limited debate as to whether adding student characteristics in addition to prior achievement scores better supports this process.
- Some research has shown that using only prior student achievement scores may be sufficient.
- Rationale for testing this assumption:
  - Statistical: To examine whether the inclusion of student characteristics reduces bias in the resulting estimates of teacher effects.
  - Policy: To examine whether the inclusion of student characteristics sets different expectations for different groups of students.

AIR

Framework for considering variables

- Data is available and accurate
- Discussion on variable inclusion
  - Is it in the teacher’s control?
  - Is it measured already by another variable?
  - Is it explained by pretest data?
- Possible definitions

Variables evaluated

Students with Disabilities (SWD) status

- Dichotomous variable for each exceptionality
- Exceptionality codes D, E, Z, U, T, M, C, F, L are excluded

Gifted status

- Dichotomous variable
- Exceptionality code L is included

English Language Learner (ELL) status

- Dichotomous variable
- Students classified as LY for two years or less

Attendance

- Continuous variable
- Number of days in attendance

Variables evaluated

Class size

- Continuous variable
- Count of students enrolled in the same course with the same teacher during the same period

Homogeneity of class composition

- Continuous variable of homogeneity of the prior-year test scores for the students within each unique course.
- For each unique district/school/teacher/course/period, for students enrolled, calculate the inter-quartile range (difference between the 75^{th} and the 25^{th} percentile score) of student test scores the prior year

AIR

Variables evaluated

Mobility

- Continuous variable
- Number of transitions between schools
- Students with one record in the current school year has 0 transitions
- Each change of school, within the year, count one transition
- If a student has two entry dates for the same school, count as one transition only if the second entry date is more that 21 days after the previous withdrawal date

Age

- Continuous variable
- Difference between student age and modal age in grade as of September 1 of the academic year

How is growth understood?

Deviation from expectation

- Given prior scores and other characteristics of the student, what is the average score of similar students? Roughly speaking, this is the expectation.
- What score did the student actually get? This is the *deviation from the expectation*, which is aggregated to comprise estimates of teacher effects.

Expected growth may be understood as the expected score minus the prior score

- Note that we lean heavily on our belief that the FCAT scale has equal intervals along the range.

Growth expectation rules

The expectation is one year's growth for the student for each course and all teachers are wholly accountable for the growth of their students.

Students in more than one course will have higher growth expectations and all the student's teachers are wholly accountable for the higher growth expectation.

AIR®

Growth expectation rules

- Students enrolled in the same course in multiple periods with the same teacher are treated as a single student in a single course
- Students enrolled in different courses with same teacher, the growth expectation is based on the number of courses and 100 percent attribution is made to the teacher for each course
- Students enrolled in different courses with different teachers, the growth expectations is based on the number of courses and 100 percent attribution is made to each teacher for each course
- Students taking the same course under multiple teachers will have the growth expectation for one course and 100 percent attribution is made to all teachers
- SWD teacher with support teacher will have the growth expectation for one course and 100 percent attribution made to all teachers

Explaining the expected scores

Model	Where the expectation comes from
Differ-ence	• Average student score among students whose prior score fell in the same FCAT Achievement Level
Model 2, 2a, 2b	• Average current score among students with similar prior scores (one or two years), similar SWD, ELL, and attendance status • All school-level effects are understood to result from the teachers
Model 3a, 3b	• Average student score among students with similar prior scores (one or two years) • School-level effects are estimated separately, and the model itself makes not commitment about attributing them to the teachers or separately
Model 3c	• Average student score among students with similar prior scores (one or two years), similar SWD, ELL, attendance and gifted status School-level effects are estimated separately, and the model itself makes not commitment about attributing them to the teachers or separately
Model 3d	• Average student score among students with similar prior scores (one or two years), similar SWD, ELL, attendance, gifted, mobility, class size, age, homogeneity • …Also—attending similar size classes with similar prior-score-diversity among students (they attend classes similar in these regards) • School-level effects are estimated separately, and the model itself makes not commitment about attributing them to the teachers or separately

Roadmap (1)

- Many variants to discuss:
  - 8 models * 7 grades * 2 subjects = 112 variants
  - 2 days offer about 8 minutes to consider each variant
- Plan: Key results are consistent across grades, so…
  - Look at pattern of estimates of the magnitude of teacher effects across all grades, all models (1 graph per subject)
  - Choose a focal grade (7) to examine results in detail

Roadmap (2)

1. (Housekeeping) The estimators we are using for the models are unbiased, consistent, and yield accurate standard errors (we present some simulation results)
2. Do the models differ in the size of effects attributed to teachers?
3. How precise are the estimates they yield and what does that mean in terms of how certainly teachers are classified?
4. What are the expectations of growth established for different groups of students?
5. What is the impact of the various models on different groups of teachers?

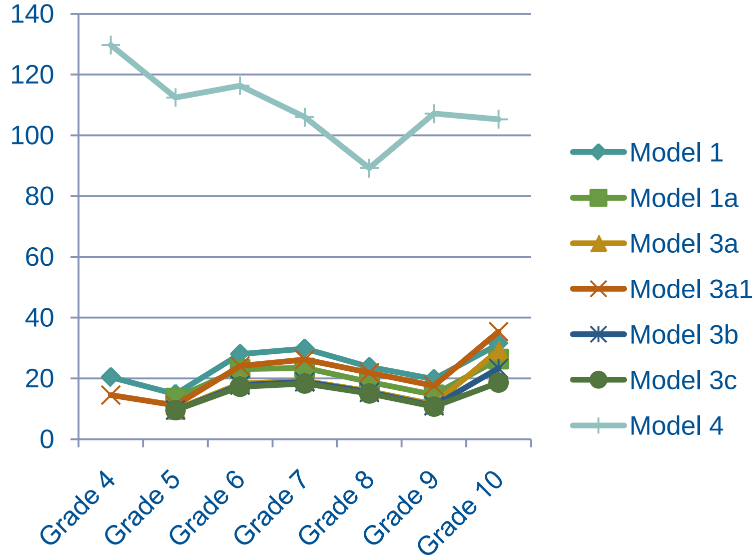
Magnitude of teacher effects

- This will be a line graph showing the magnitude of the effects, one line for each model, across grades
- A second slide will show that the school and teacher effect models are about the same when the two effects are added together
- The next pair of slides will show the same for reading

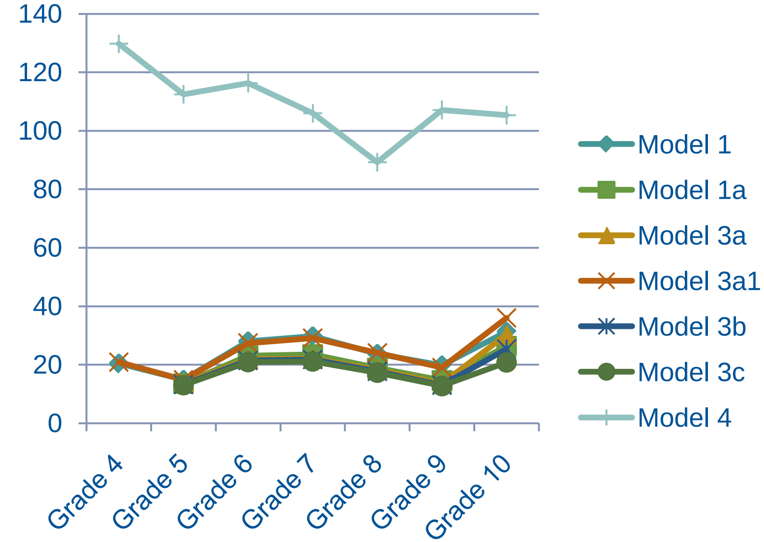
Magnitude of teacher effects: Reading

Estimated size of teacher effects in DSS scale score points

Teacher effect, when accounting for school effect



Teacher effects when all school effects are attributed to teacher



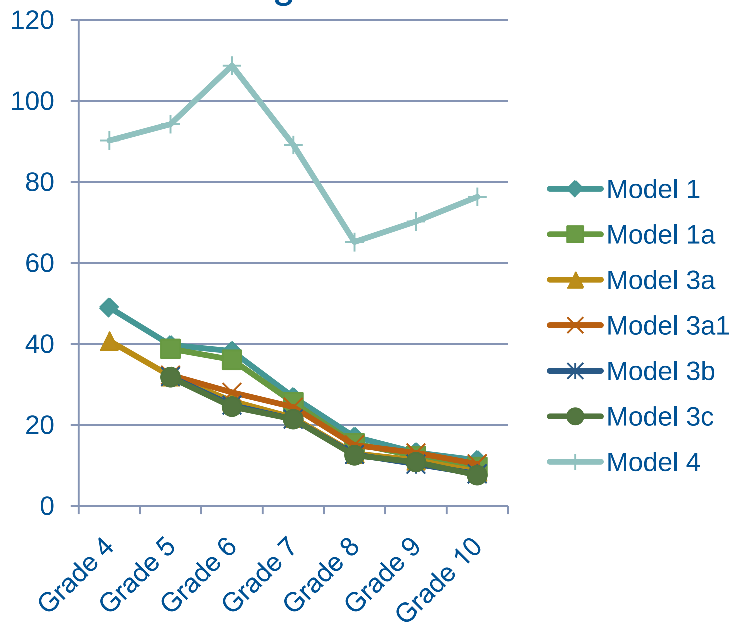
All of the models show similar patterns of teacher effects across grades. When models separate teacher and school effects, policy can determine how much of the school effect is attributed to teachers.

AIR®

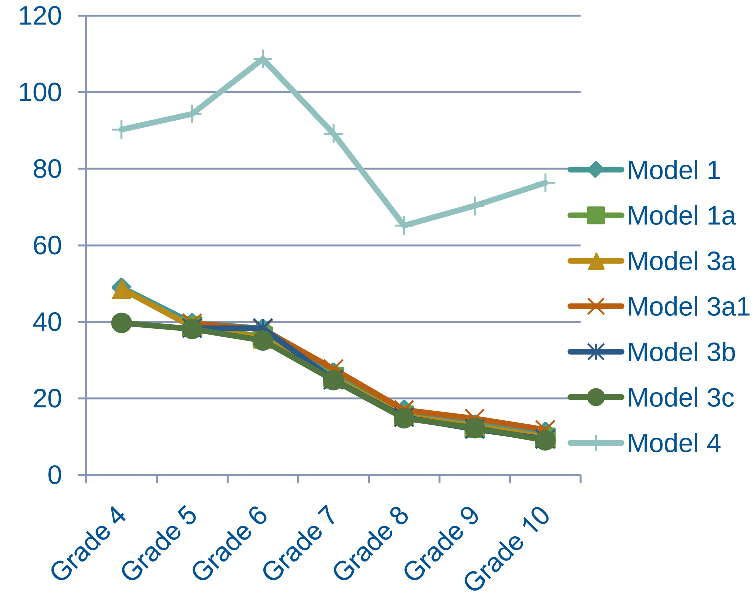
Magnitude of teacher effects: Math

Estimated size of teacher effects in DSS scale score points

Teacher effect, when accounting for school effect



Teacher effects when all school effects are attributed to teacher



The patterns are different between reading and math. This is not surprising because the two scales are not comparable. All models once again show similar patterns.

Focus on grade 7

- Recall that there are 98 grade x model x subject variants
- Results are consistent across grades (numbers vary, but inferences are the same)
- We present detailed results for grade 7, math and reading

AIR®

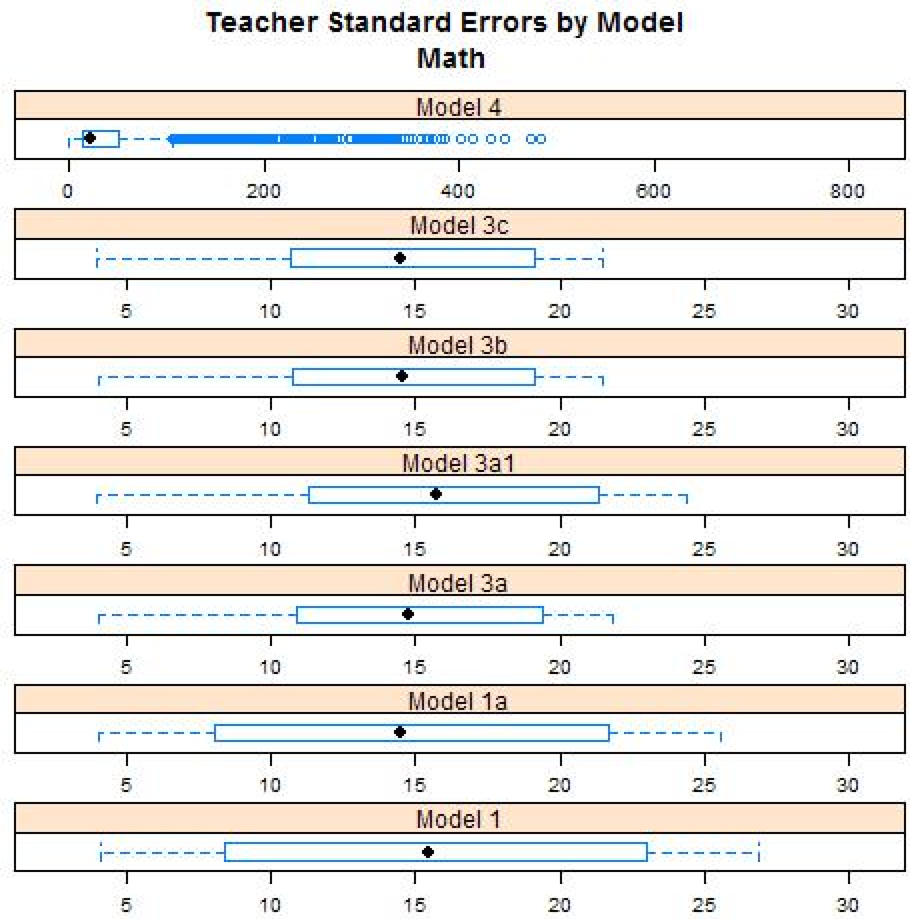
Precision of the teacher effects

- **Question**: What characteristics of value-added models lead to more **precise** estimates of the teacher effects?
- **Statistic to examine**: Standard errors of the teacher effects.
- **Evidence in favor of a desirable model** : A model with small standard errors, other things being equal, is more desirable than a model with larger standard errors.
- **Why**: A smaller standard error tells us that the estimated teacher effect is more precise under a certain model.

Precision and uncertainty: Standard errors, Math

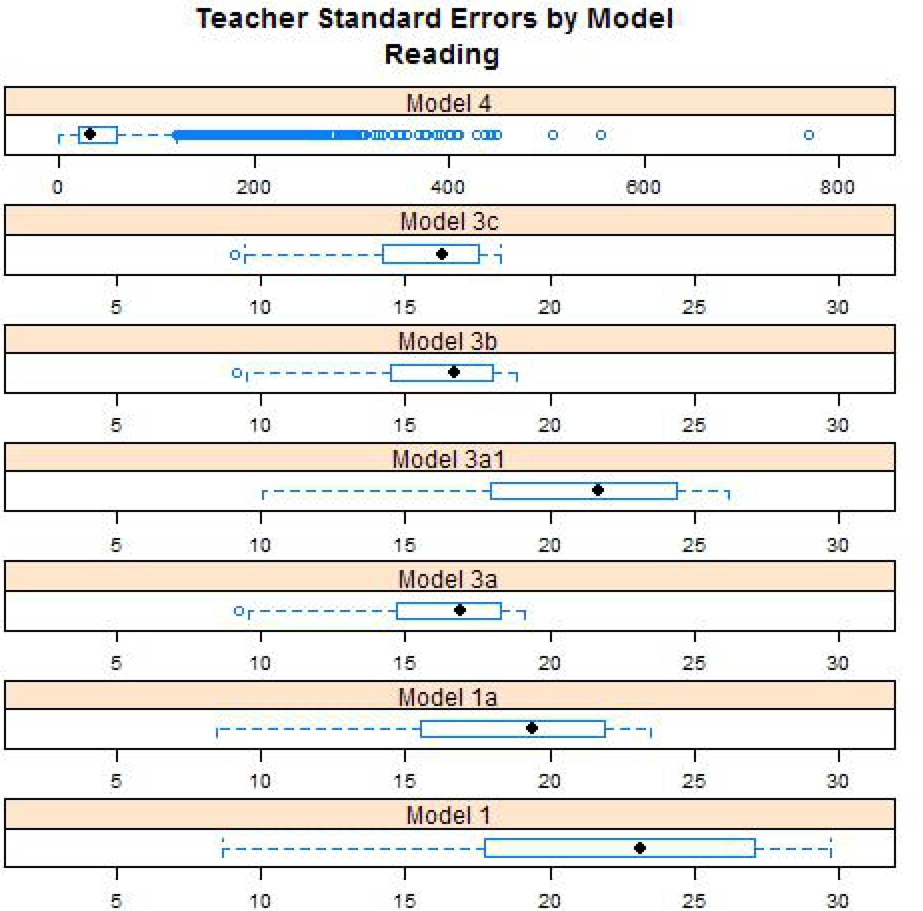
All models except the differences model have median standard errors near about 15 DSS points.

Models 1a,3a,3b, and 3c are all a bit more precise. They (and only they) include an extra prior year's data.



Precision and uncertainty: Standard errors, Reading

We again see that the models with the extra prior year data yields more precise estimates of teacher effects.



Models in order of their precision

READING: Model	Does the model include school effects?	Does the model include 2 prior scores?
3c	Yes	Yes
3b	Yes	Yes
3a	Yes	Yes
1a	No	Yes
3a1	Yes	No
1	No	No
4	No	No

MATH: Model	Does the model include school effects?	Does the model include 2 prior scores?
3c	Yes	Yes
3b	Yes	Yes
1a	No	Yes
3a	Yes	Yes
1	No	No
3a1	Yes	No
4	No	No

We again see that the models with the extra prior year data yields more precise estimates of teacher effects.

School effects or teacher effects

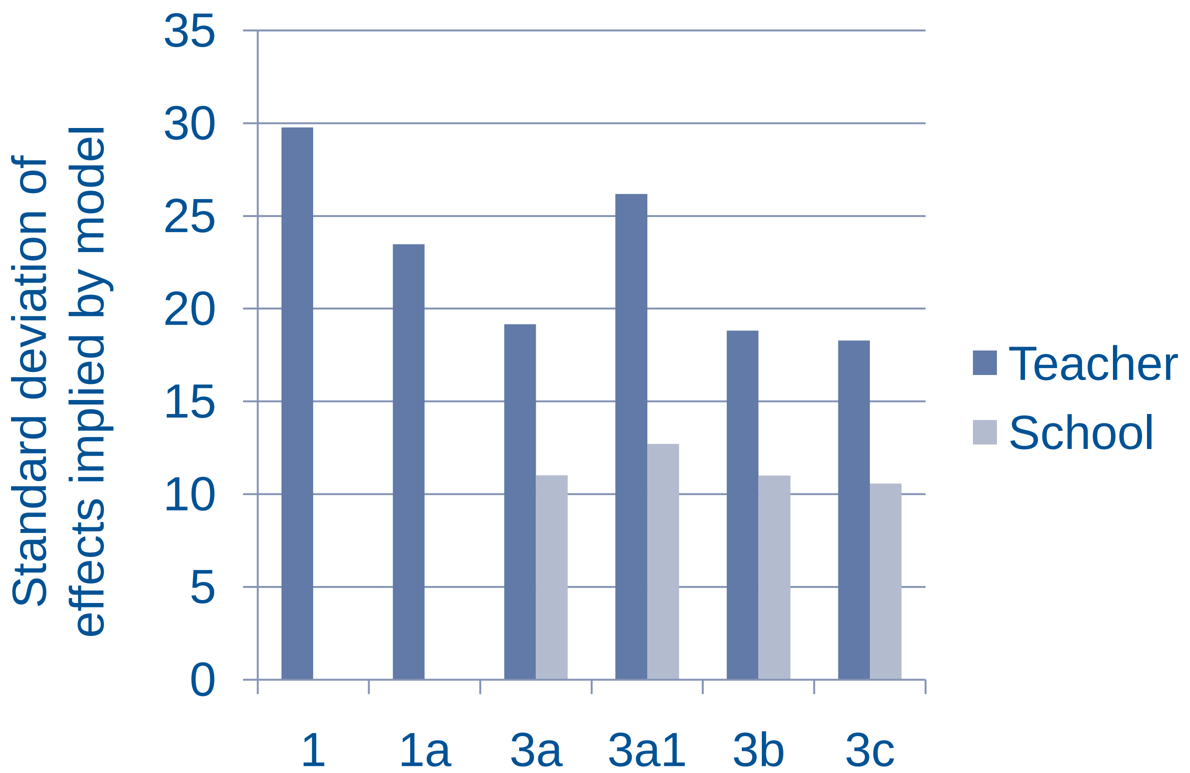
Question: Should the value added model account for systematic differences between schools?

Statistic to examine: Variation in student growth between schools

Evidence in favor of a desirable model : If models suggest that systematic school effects exist, policy must decide how much to attribute to teachers

Why: Determining if, and how much of, the school level effect should be attributed to a teacher

Variances between teacher and school (Reading)



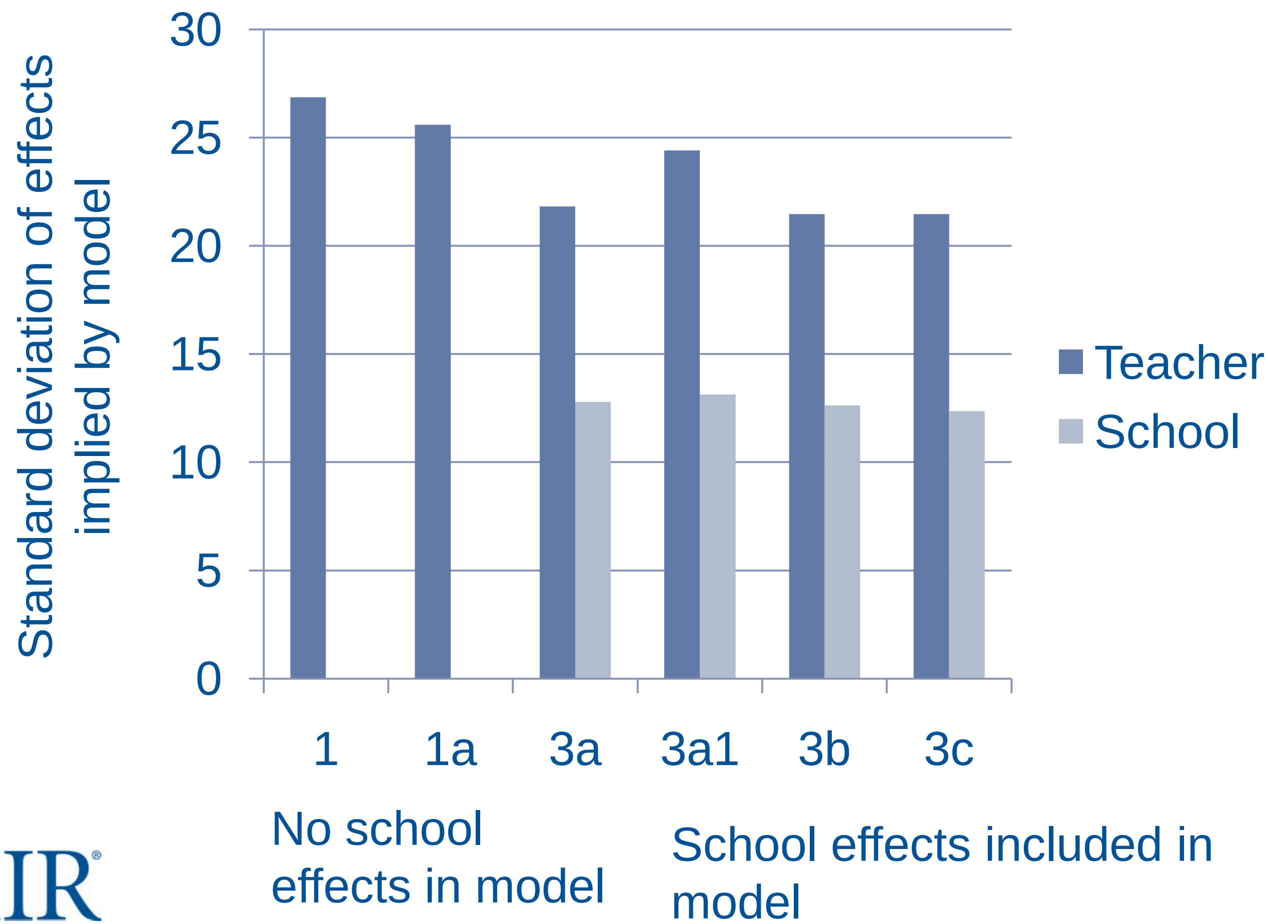
School effects appear to be real, but smaller. Models 1 and 1a implicitly assign these effects to teachers. This choice should be explicit.

No school effects in model

School effects included in model

AIR®

Variances between teacher and school (Math)



We see the same pattern in math.

Two consistent findings with school effects

- Schools appear to account for some non-trivial variation
- Teachers appear to be less different from each other when including school effects
- School effects may be entirely attributable to teachers, or they may result from other factors. Also…could be both.

AIR

Model parsimony

Question: Does the model include control variables without being overly complicated?

Statistic to examine: Percent of current year test score variance accounted for by control variables in models

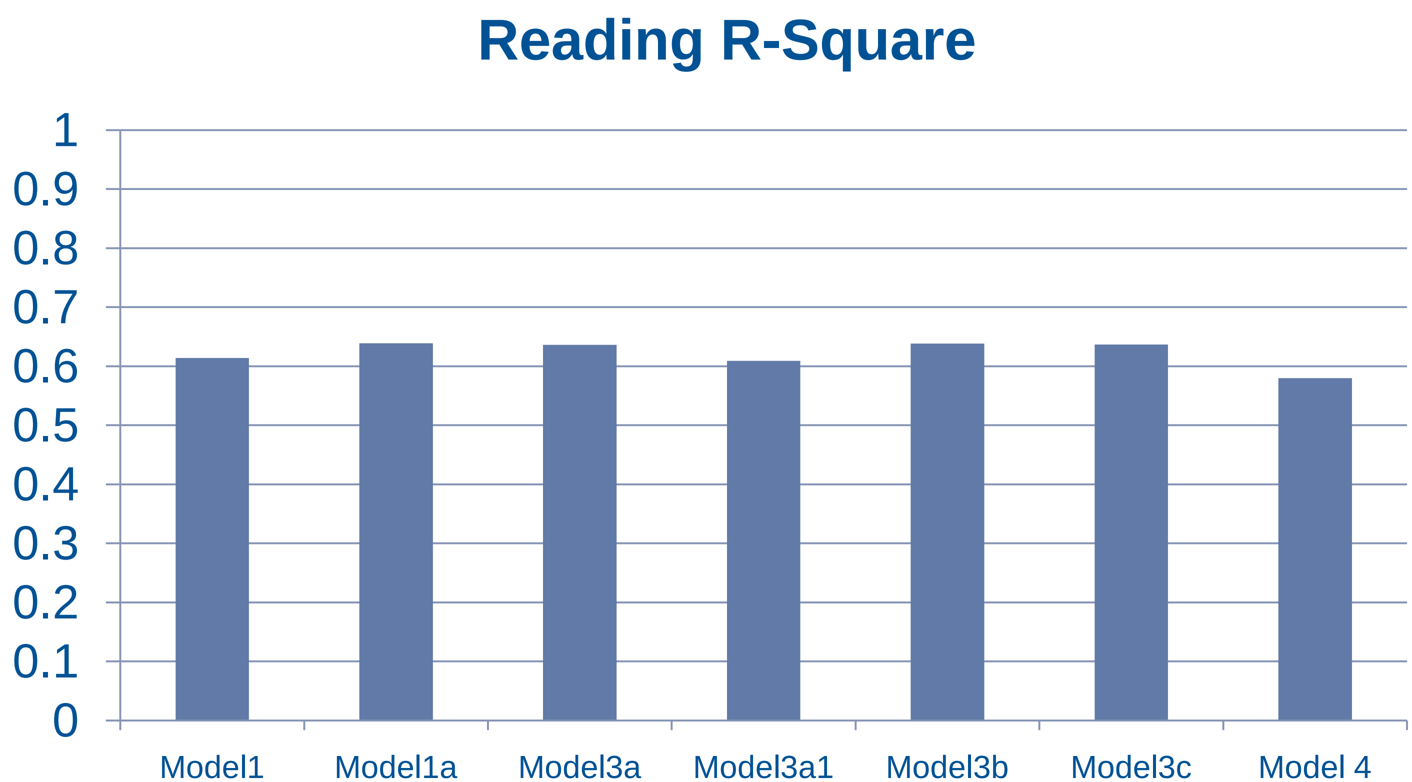
Evidence in favor of a desirable model: High proportion of variance accounted for

Why: The model should not be needlessly complex

Proportion of variance in current year test score explained by control variables (Reading)

More control variables improves model fit.

An additional prior year gives a moderate difference, other variables matter less



SWD, ELL, Attendance, Gifted

1 prior   2 prior

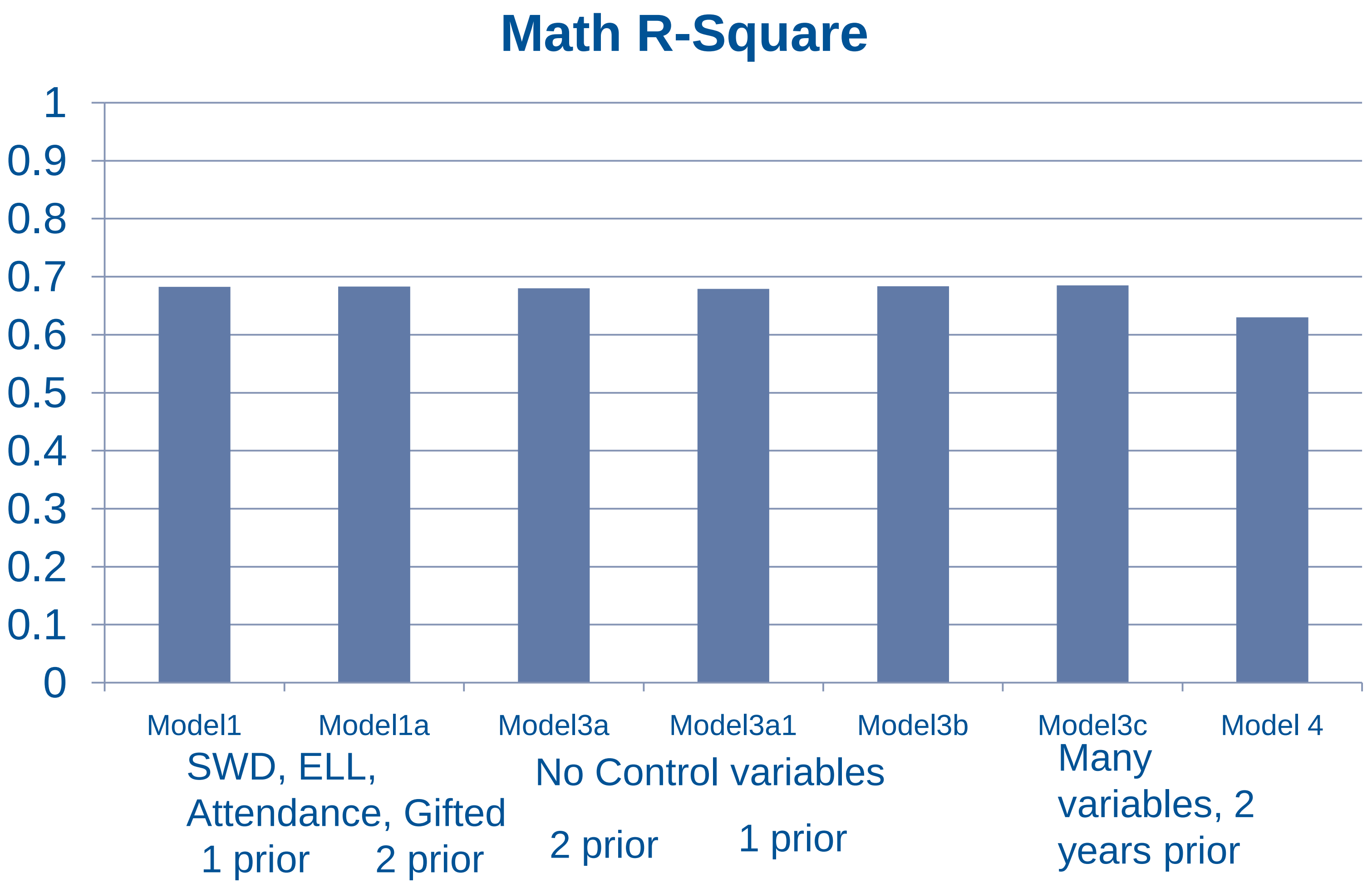
No Control variables

2 prior   1 prior

Many variables, 2 years prior

Proportion of variance in current year test score explained by control variables (Math)

We see a similar pattern in math



What can we learn about parsimony?

- Models 3b and 3c have very similar R-Square values in reading and math
- This suggests that the additional variables in 3c do not help form better estimates than what is obtained in Model 3b

AIR®

Control variables

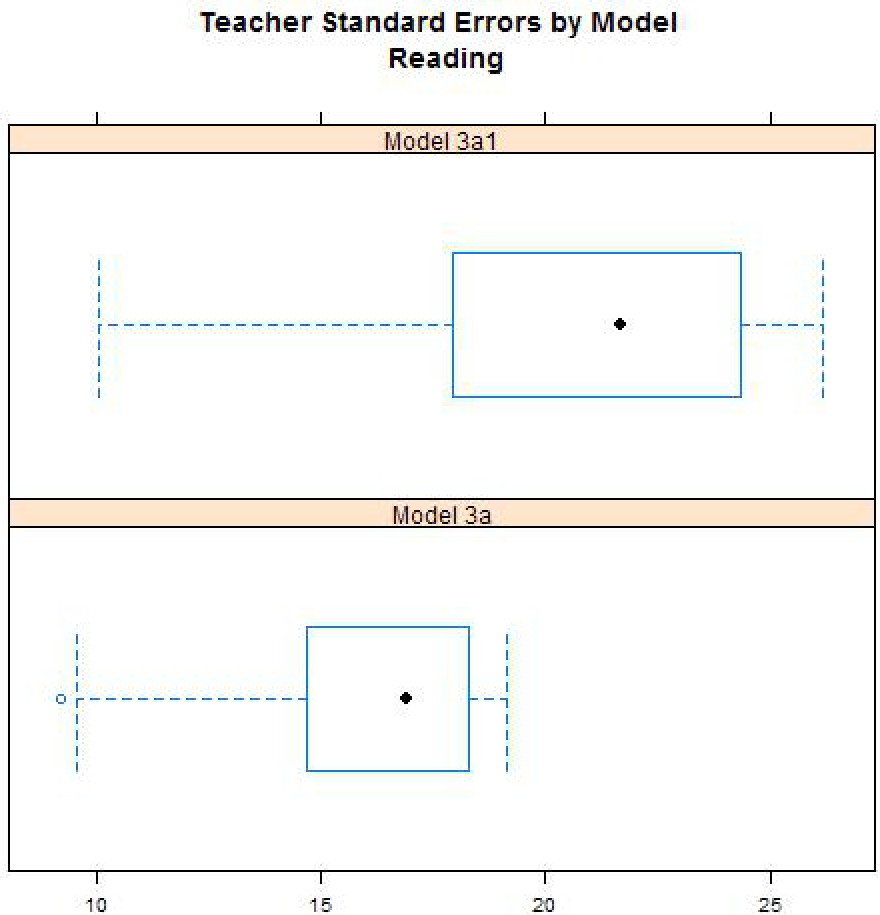
Additional control variables (excluding technical necessities and prior achievement)	Model 3a (2 lags, no control variables)	Model 3b (2 lags, SWD, ELL, Attendance	Model 3c (2 lags, all regressors)
Total variance explained			
SWD 3- Language impaired		-5.82	-2.01
SWD 4- Hearing impaired		6.65	9.62
SWD 5- Visually impaired		-10.08	-8.85
SWD 6- Emotional/behavioral		1.16	2.82
SWD 7- Specific learning disability		-2.07	2.05
SWD 9- Dual sensory impaired		-121.63	-129.03
SWD 10- Autism spectrum disorder		11.42	12.55
SWD 12- Traumatic brain injury		-31.55	-27.56
SWD 13- Other health impaired		-7.92	-5.36
SWD 14- Intellectual disability		-13.71	-8.36
Class 1 size			-0.21
Class 1 homogeneity			0.00
Class 2 size			-0.12
Class 2 homogeneity			0.03
Class 3-6, size and homogeneity			NS
Difference from modal age			-7.82
Mobility			-5.40
Attendance		0.18	0.16
_0910_S_Gifted		-0.31	-0.40
_0910_ELL_LY		28.96	28.79

Include 1 or 2 years of prior achievement data

- **Question**: Should the value added model include 1 or 2 prior achievement data for each student
- **Statistic**: Standard errors
- **Evidence in favor of a desirable model**: Lower standard errors
- **Why**: More prior information about students may provide better insight into their expected growth and to help better estimate a teacher effect

Does more prior achievement data improve estimates (Reading)

Adding an extra year of prior achievement yields more precise estimates

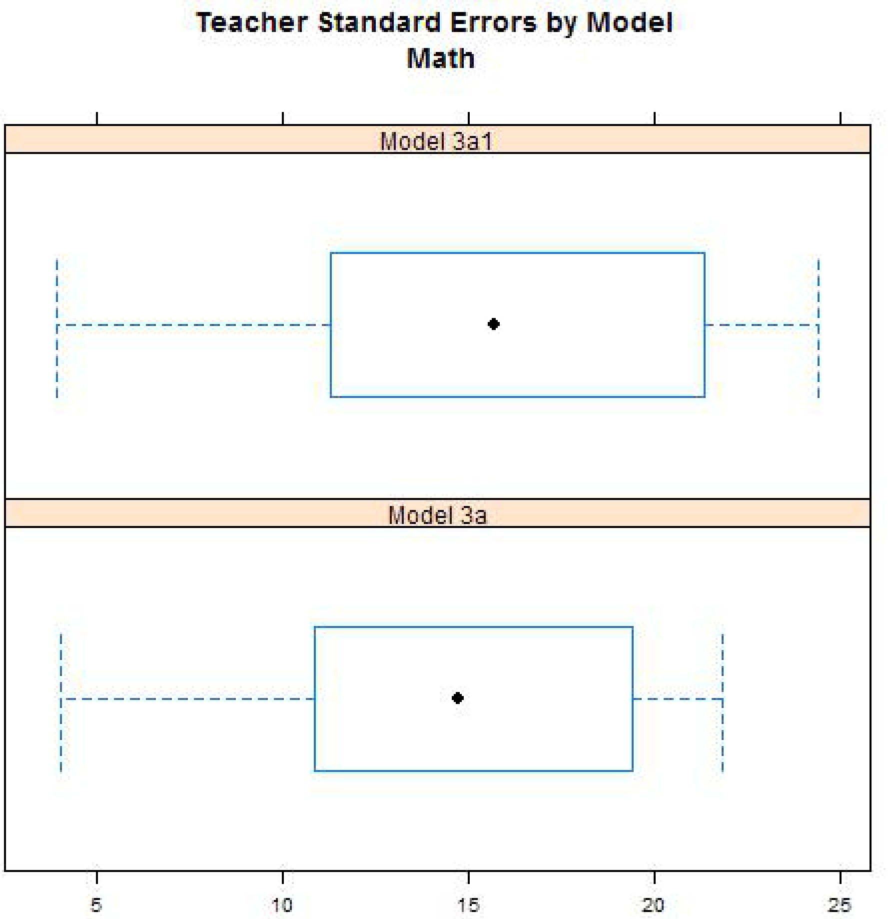


This model includes only one year prior achievement and nothing else

This is the same model with two years prior achievement

Does more prior achievement data improve estimates (Math)

We see the same pattern in math



This model includes only one year prior achievement and nothing else

This is the same model with two years prior achievement

Rank models by number of prior years of achievement data

Of the two comparable models, standard errors are smaller when two prior years of achievement are included

Tentative model selection

Which model looks best so far?

The next step is to examine the implications of the model:

- What does it imply about expectations for students?
- Which teachers get higher value added score?

Let’s take a moment to see where we are…

Model Characteristic	Committee Judgment
Differences or covariate model?	?
Should it have school effects?	?
One or two years prior achievement?	?
None, many or few control variables?	?

Expectations

- Pairs of bar charts, math and reading for the following variables
  - SWD classifications by model
  - ELL by model
  - Gifted by model

- Pair of scatter plots
  - Prior expectation by prior achievement

Questions from Thursday

- Impact of school effects and attribution to the teacher
- Examples of teacher value added scores under the different models
- Cell size, discussion

Questions from Thursday

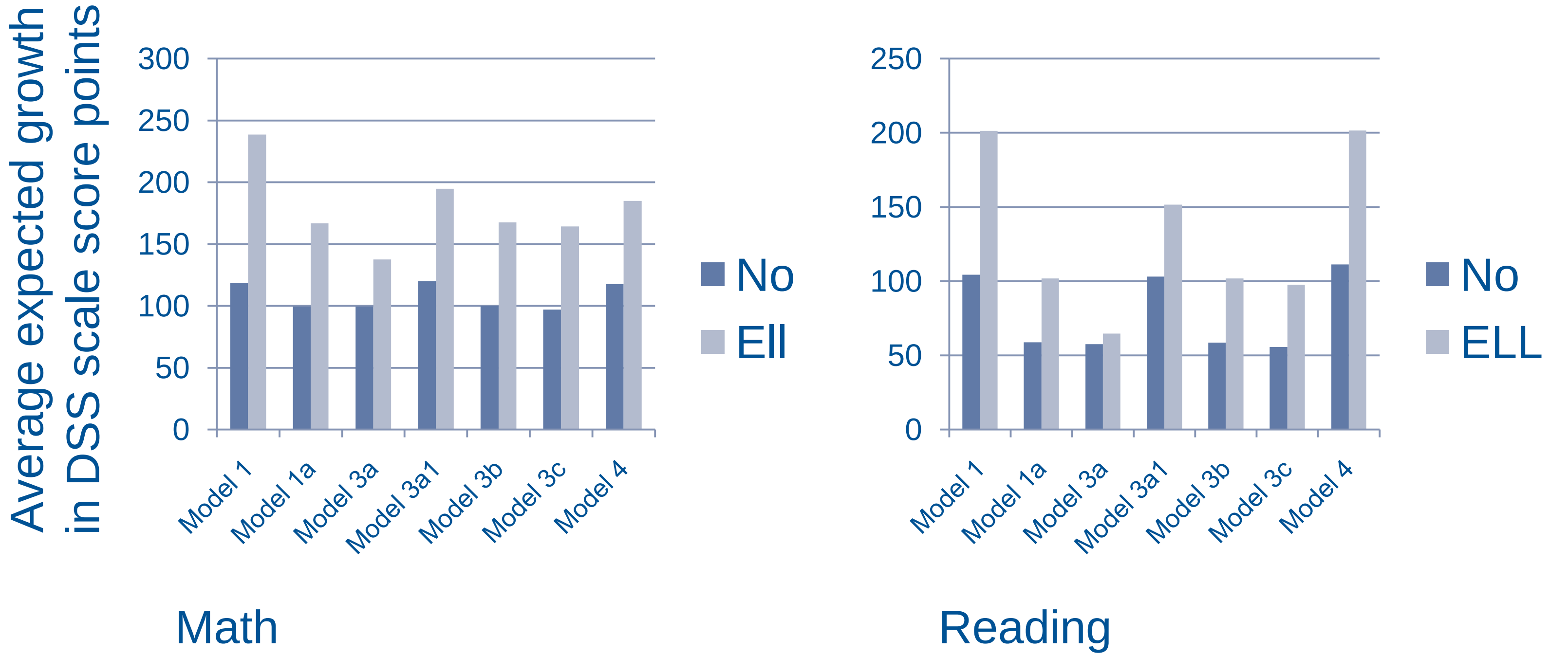
- Regression coefficients for reading and math and all grades for model 3c
- List of variables that were not significant for any subject and grade level, and what was the greatest effect size of the non-significant variables

AIR®

What do the models imply about growth expected from different students?

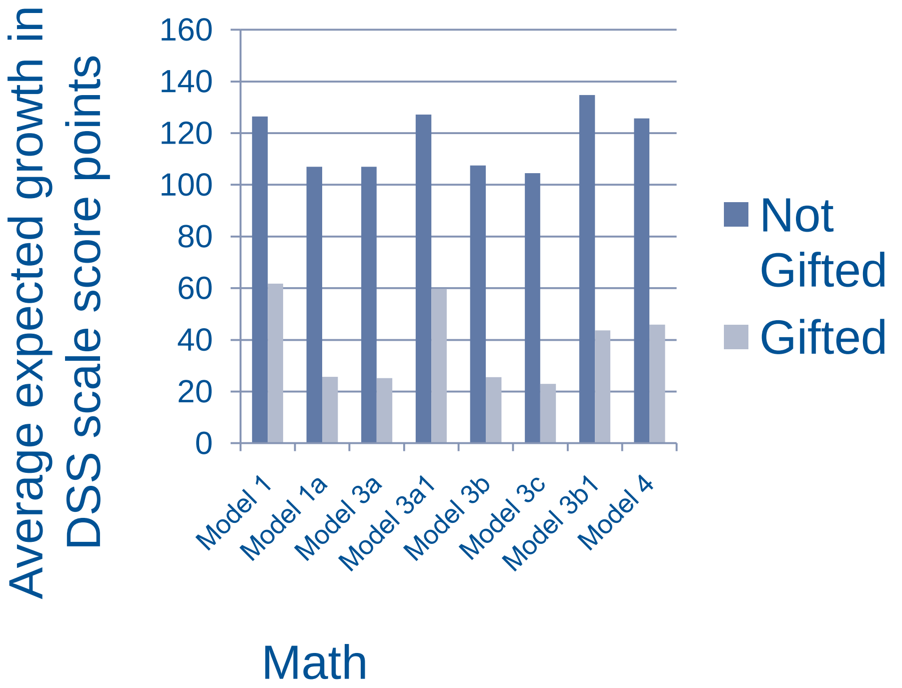
- Recall: These models measure effectiveness as the deviation from expected performance
- All the models display a negative relationship between expected growth and prior performance
  - Students with lower prior performance typically gain more DSS points in a year than higher achieving students
    - May be a measurement artifact

Expectations for ELL students



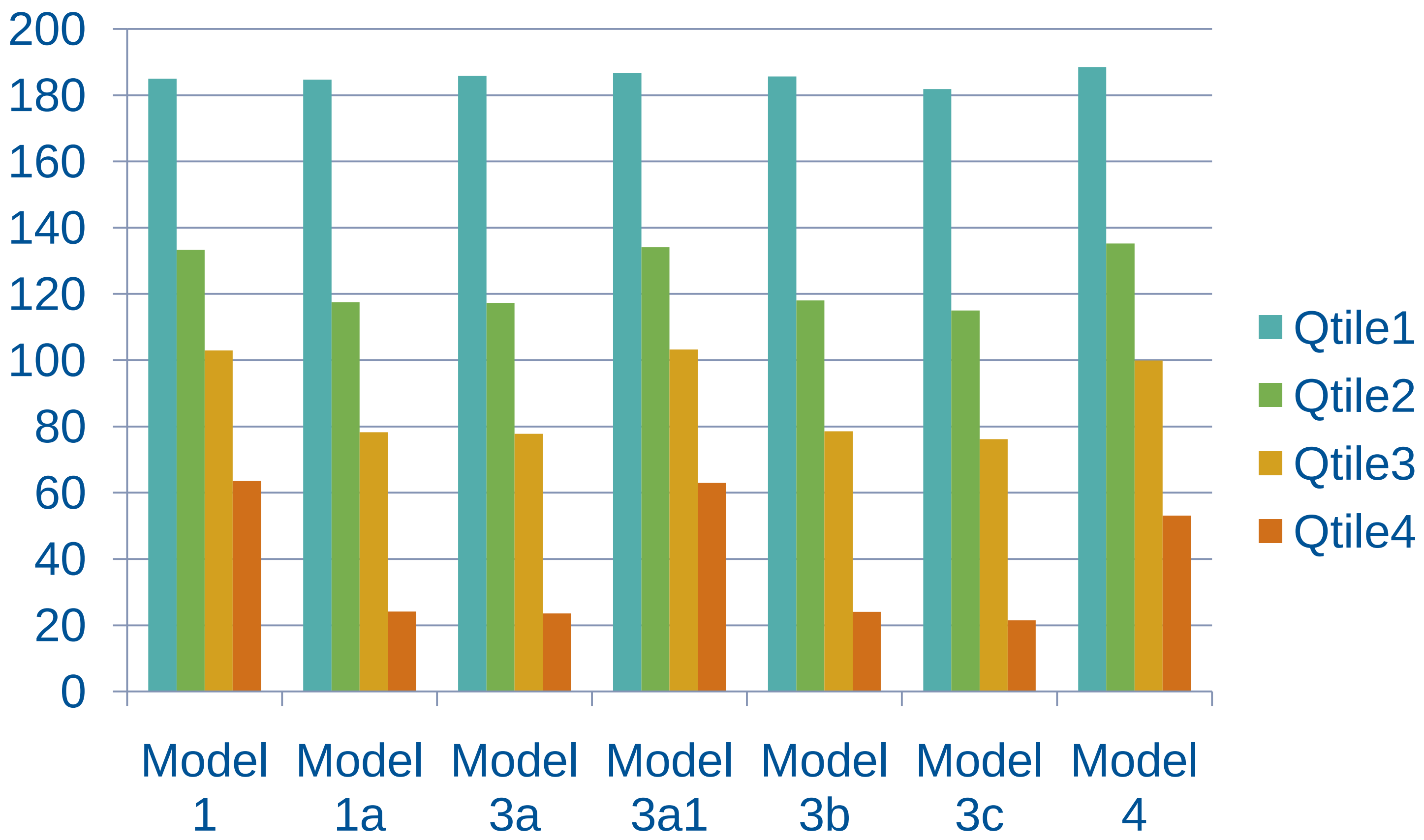
ELL students typically traverse more scale score points than non ELL students in a year in all models.

Expectations for Gifted students

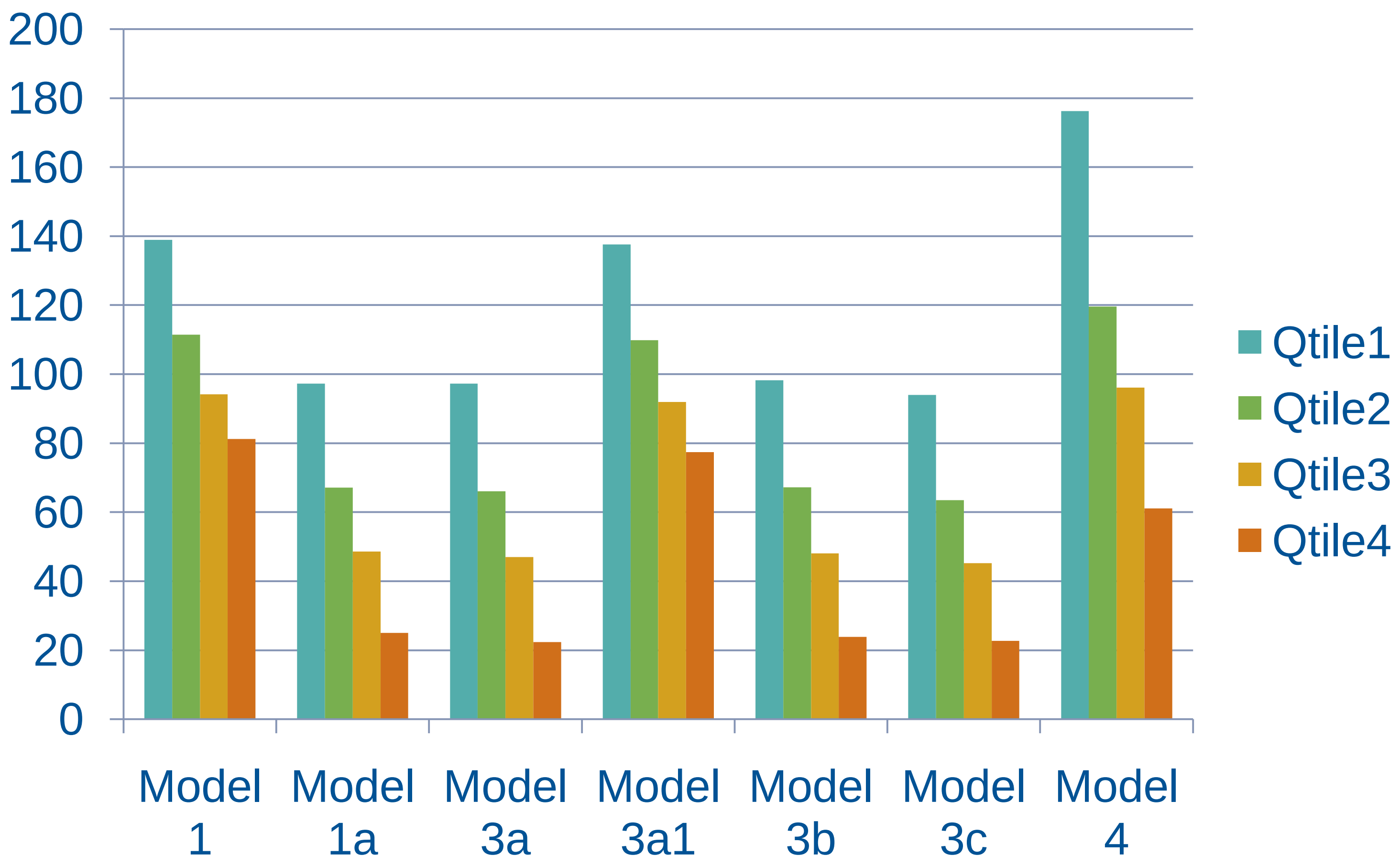


Non-Gifted students tend to have larger growth expectations than Gifted students

Expectations by prior DSS performance quartile (Math)



Expectations by prior DSS performance quartile (Reading)

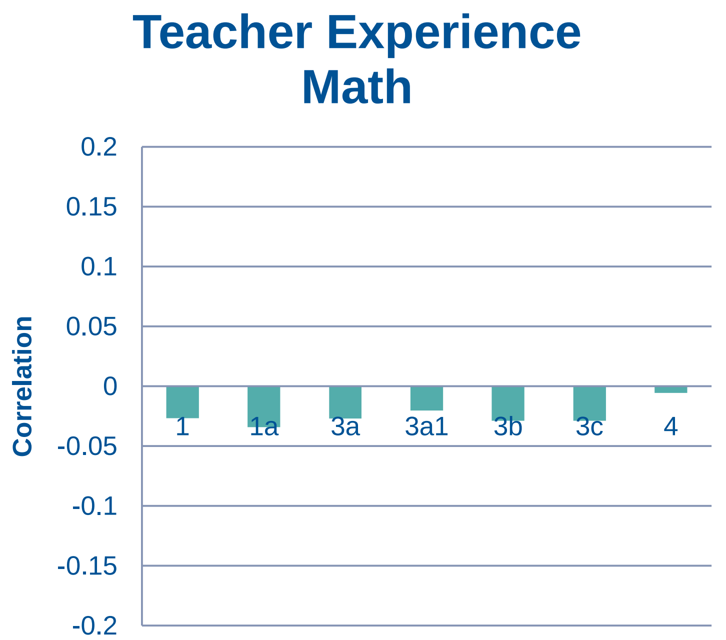


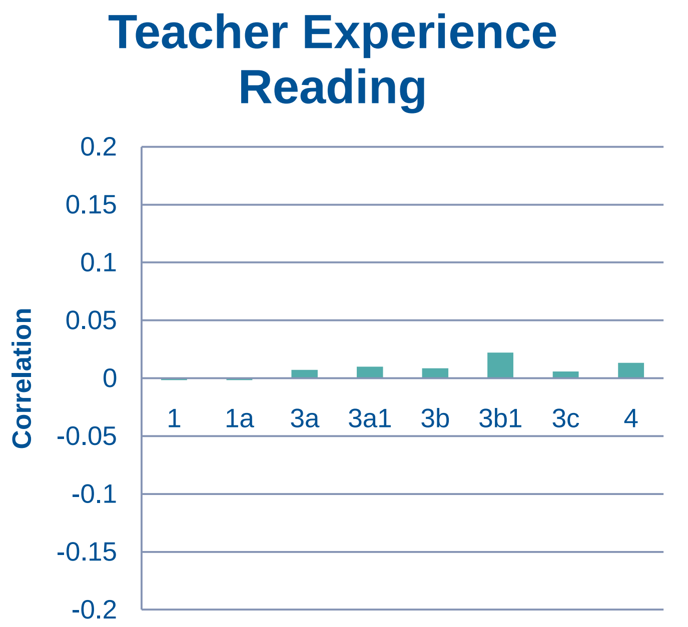
Impact: Which teachers score above average

We would expect that a good value-added measure would be associated with things that we believe are associated with good teachers, and not associated with other things.

We will look at these relationships.

Correlation between teacher experience and value added measures

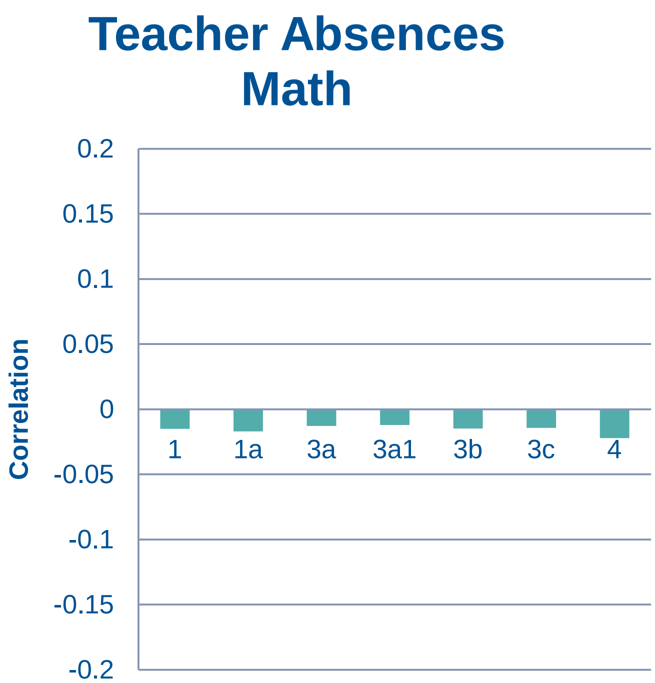


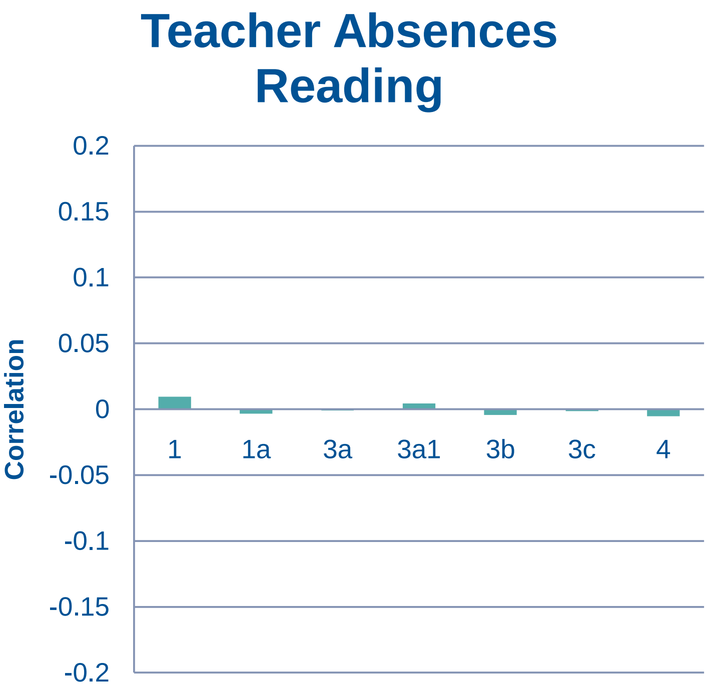


None of the model results are correlated with teacher experience. In part, we suspect that is due to inconsistent or inaccurate teacher experience data.

AIR®

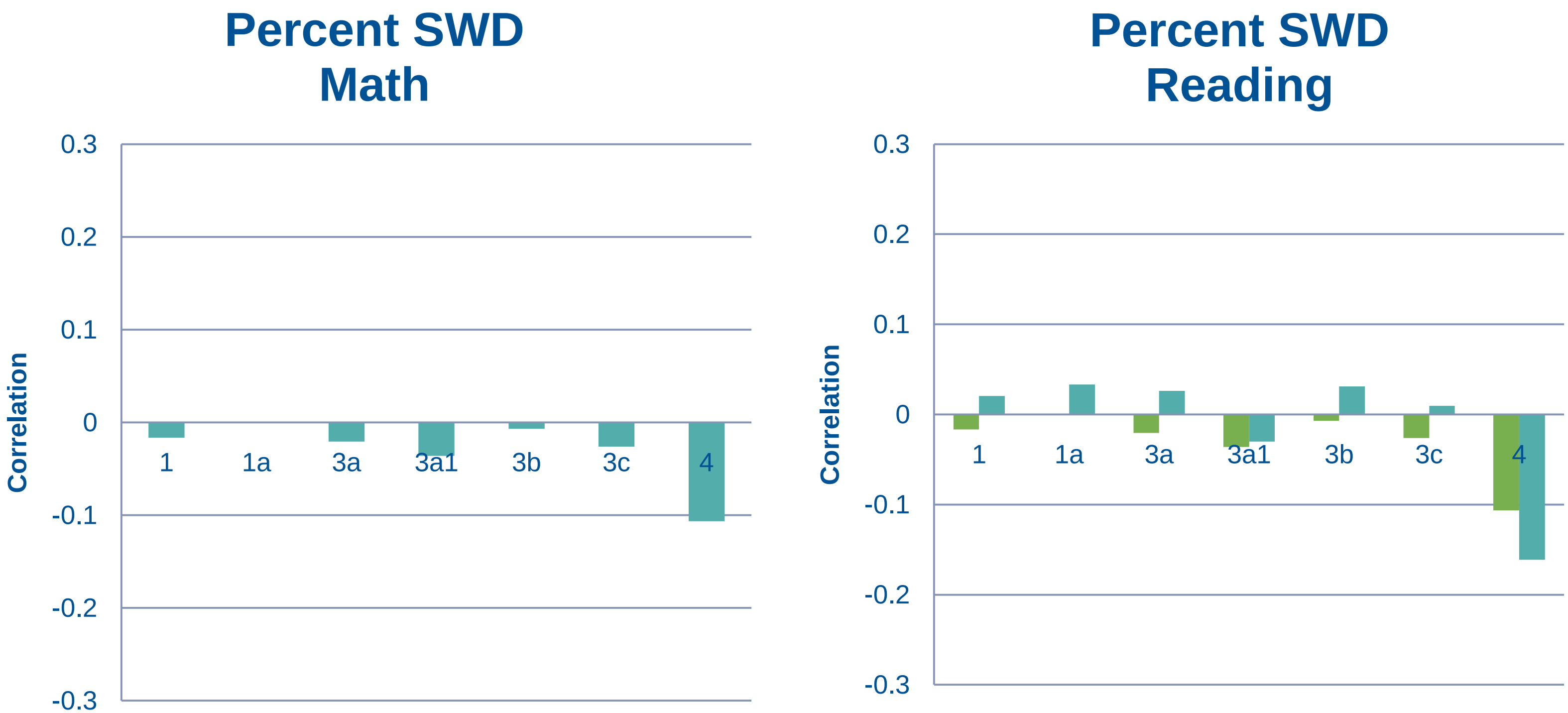
Correlation between teacher attendance and value added measures





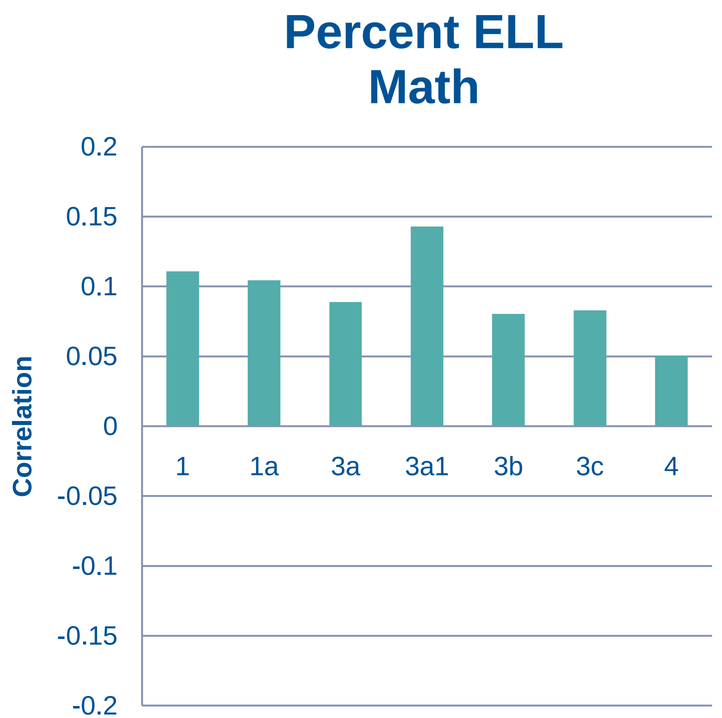
None of the model results are correlated with teacher attendance.

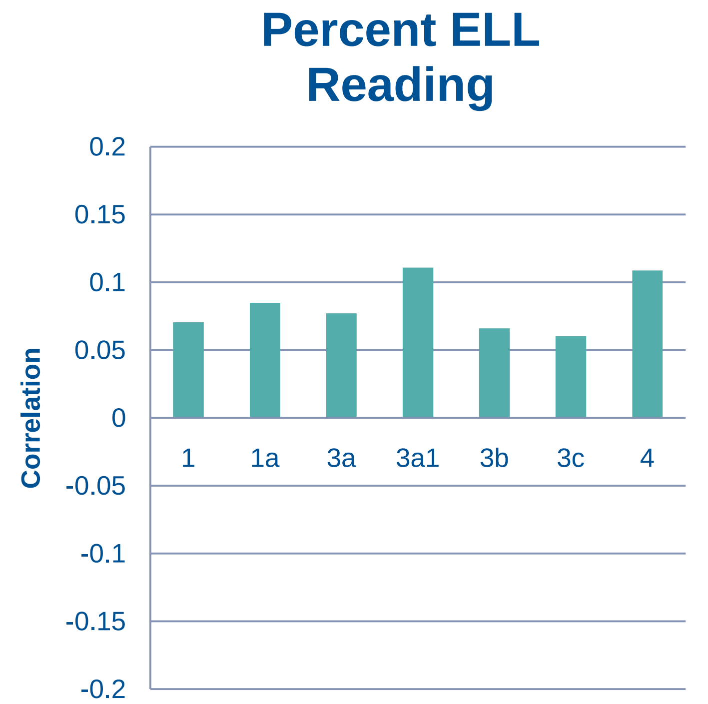
Percent SWD taught correlation with value added measures



Only Model 4, the difference model, is correlated with the percent SWD in the class. Teachers with many SWD students under this model are slightly more likely to have lower value-added measures.

Percent of ELL students taught correlation with value added measures

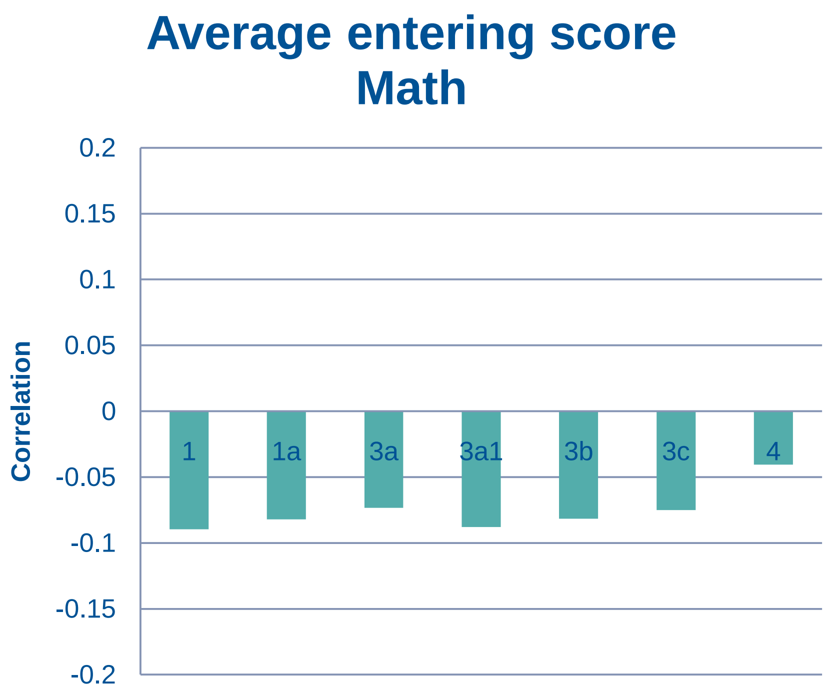


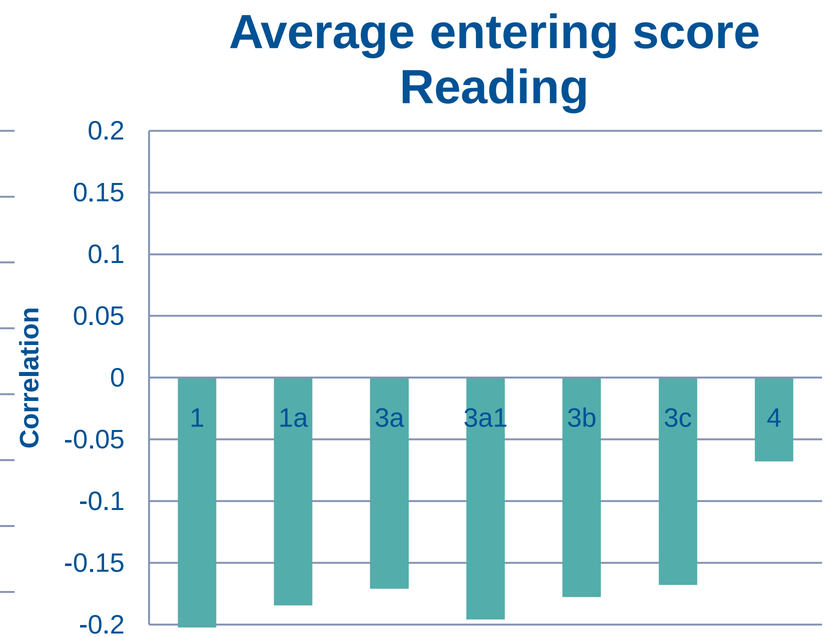


In all the models, teachers teaching more ELL students appear receive slightly higher value-added scores—even when ELL is included as a control variable.

AIR®

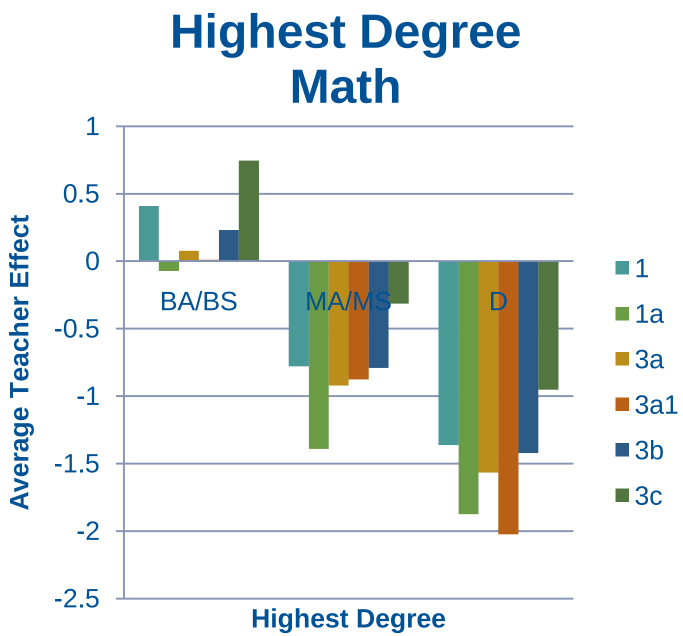
Average entering score correlation with value added measures

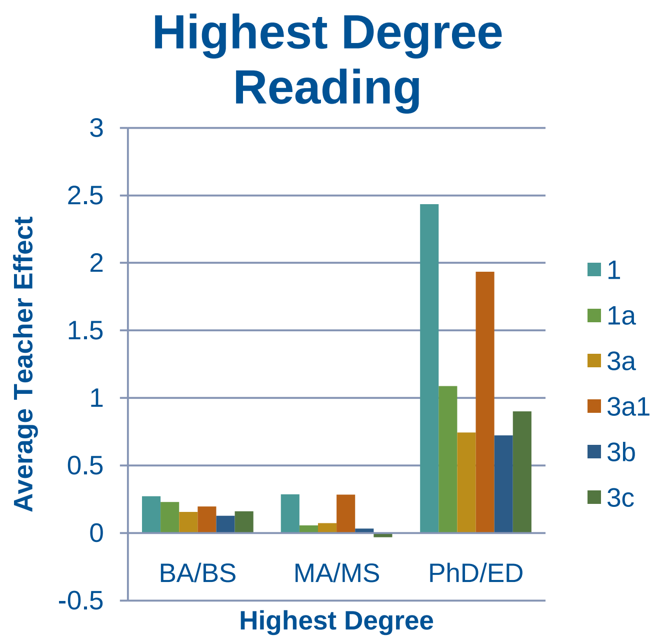




All of the value-added measures have a small to moderate negative correlation with the average score of students entering the class. This implies slightly higher value-added scores among teachers of lower-performing students.

Value-added scores by teacher degree





Teachers' degrees show an odd pattern of relationship. In reading, a Master's degree seems to confer no advantage, but a PhD/EdD does. In math, graduate degrees are associated with lower value-added scores.

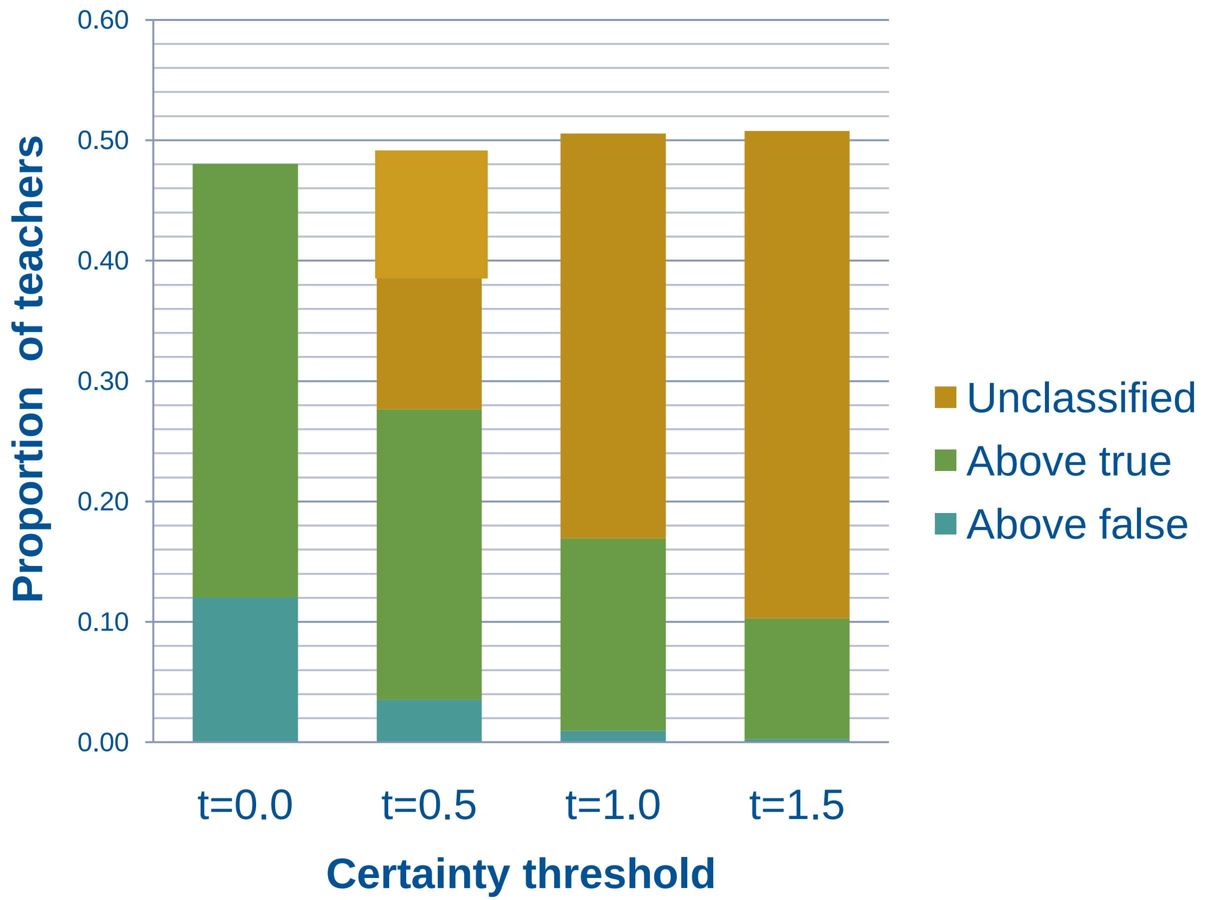
AIR®

Classification

- Once scores are in hand, they can be used in a variety of ways
- Classification decision can invite or avoid misclassification
- This section highlights some risks, and suggests some mitigation strategies

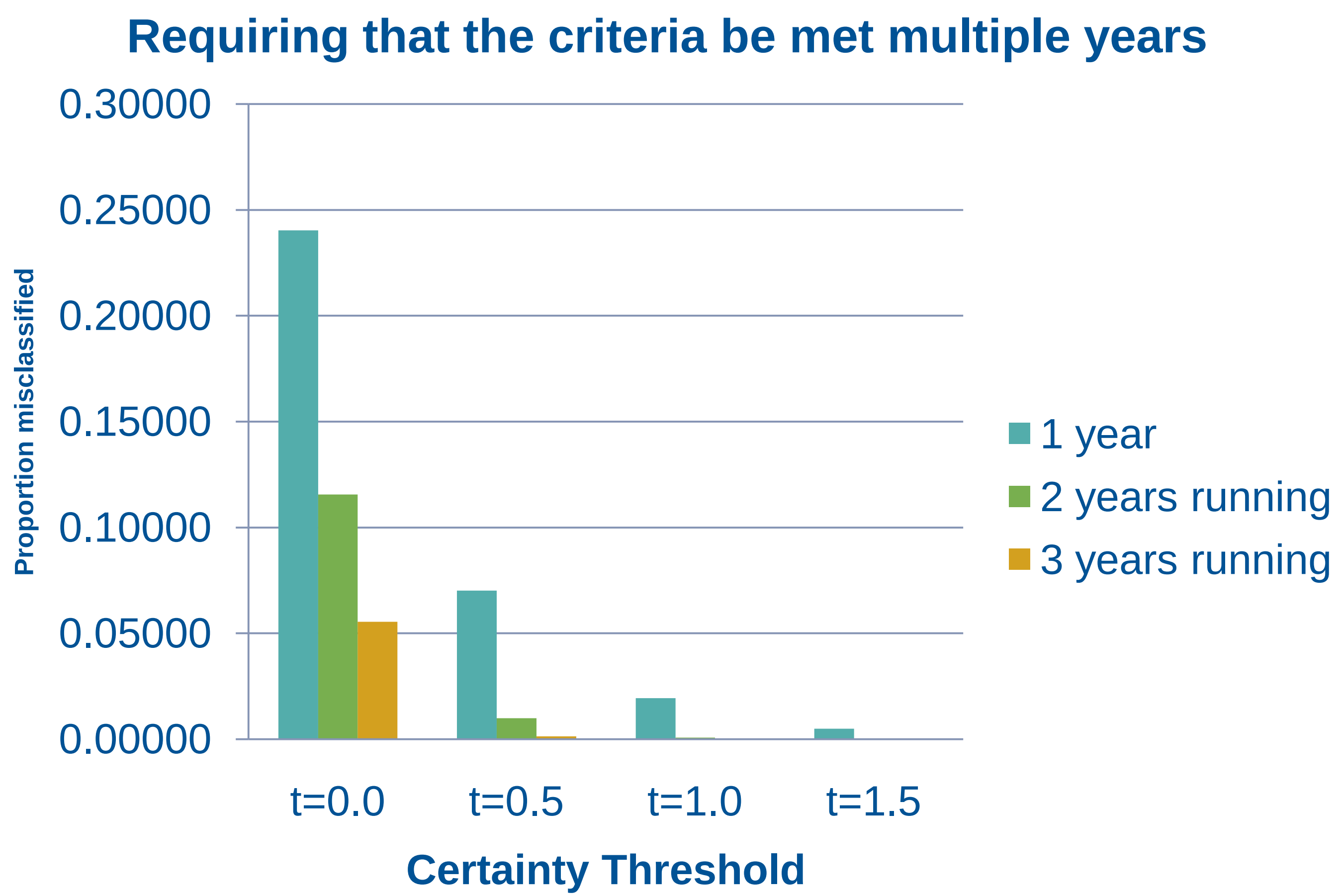
Policy decisions can dramatically reduce misclassifications

Adding the requirement that teachers be above the mean with some certainty (at least 0.5, at least 1, at least 1.5 SE) reduces the misclassifications, but also reduces the number of teachers classified



Adding the requirement for multiple years

When the requirement must be met multiple years in a row further decreases misclassification



SGIC Recommendation

Model

Variables

Identification of courses by FLDOE Course Code Directory (CCD)

Reading

- require a reading endorsement and/or
- used for reading intervention

English/Language Arts

- identified based on CCD prefixes
  - 5010 for grades K-5 and 10 for grades 6-12
- reviewed by statewide committees

Math

- identified based on the CCD prefixes
  - 12 for K-5 and 50 for grades 6-12
- reviewed by content specialists to ensure the listing was complete

Identified courses in 2009-10 Course Code Directory (CCD)

Reading and English/Language Arts

- 166 courses
- Reading, Debate, Speech, Screen Play Writing, English, Communications, Creative Writing, Literature, Mass Media, Journalism, Great Books

Math

- 90 courses
- Pre-Algebra, Algebra, Academic Skills K-5/6-8, Life Skills Math 9-12, Trigonometry, Discrete Mathematics, Consumer Mathematics, Geometry, Calculus. Math for College Success

AIR

Proposed process for identifying courses for value added

- Brief FLDOE content staff on the SGIC value added work
- Request FLDOE content staff make recommendations to SGIC on which reading and math courses are aligned to FCAT for use in teacher evaluation
- SGIC meet mid-June to review recommendations of FLDOE content staff and propose course inclusion

White Paper Outline

Overview of SGIC Meetings

Meeting	Date	Topics
Webinar	March 24, 2011	Introductions, project and process overview
In Person Orlando	April 4-5, 2011	Overview of value-added models; eight different types to analyze; discussion of business rules; selection of factors; direction from committee on which models to review
Webinar	April 14, 2011	Variables selection
In Person Orlando	May 19-20, 2011	Present and discuss results of analysis of the eight different models and form preliminary recommendations on final model
Webinar	**May 25, 2011, 4:30–6:30 pm**	**Reach consensus on recommendation for the final model to present to the Commissioner on June 1**
Webinar?	Mid-June	Review, discuss, recommend course inclusion for statewide FCAT value added models

AIR®

Questions and Next Steps

Information about the activities, membership, meeting schedule and materials, and recording of conference calls and webinar of the SGIC are posted at: www.fldoe.org/arra/racetothetop.asp.

AMERICAN INSTITUTES FOR RESEARCH®

Contact Information

FLDOE:

Juan Copa, Director of Research and Analysis in Educator Performance

850-245-0744 (office)

Juan.Copa@fldoe.org

AIR:

Christy Hovanetz, Ph.D., Project Director

850-212-0243 (cell)

ChristyHovanetz@gmail.com