Evaluating End-of-Course (EOC) Models for Educator Evaluation

Presentation to Student Growth Implementation Committee

December 10, 2013

AIR®
American Institutes for Research®

Agenda

- 8:30 — Welcome, Agenda Overview: Juan Copa, *FLDOE, Deputy Commissioner*

  Updates: Ronda Bourn, *Chair, SGIC*
- 9:00 — Stability Analysis for FCAT Model: Harold Doran, *AIR*
- 9:15 — End of Course Model Options: Harold Doran, Eric Larsen, and Dan Sherman, *AIR*
- 10:30 — Break
- 10:45 — End of Course Models Continued
- 12:00 — Lunch on your own

Agenda

- 1:15 — Discussion of End of Course Models: Ronda Bourn and Juan Copa
  Discussion of District Approaches
- 3:00 — Break
- 3:15 — Recommendations for End of Course Models: Ronda Bourn and Juan Copa
- 4:00 — Next Steps: Juan Copa
- 4:30 — Adjourn

Stability Analysis for FCAT

AIR®
American Institutes for Research®

Stability of FCAT Teacher VAM Scores

- How much does a teacher's VAM score change from year to year?
- Reasons a teacher's VAM score changes between years:
  - Changes in the teacher's effectiveness
  - Year-to-year "noise" in the estimates
- AIR calculated the following correlations:
  - Between 2012-13 and 2011-12 teacher VAM scores ("one-year scores")
  - Between 2012-13 and aggregate 2011-12, 2010-11, and 2009-10 VAM scores ("three-year scores")
  - Separate correlations for reading, math, and combined scores

Correlations Between Teacher VAM Scores Over Time

	2011-12 One-Year Scores			2011-12 Three-Year Scores		
	Reading	Math	Combined	Reading	Math	Combined
2012-13 One-Year Score	0.28	0.46	0.38	0.25	0.47	0.35

Stability of FCAT Teacher VAM Scores

- These correlations are similar to those found in other research
- Correlations in math higher than correlations in reading
  - Because teacher value-added estimates are more precise in math, this is what we would expect
- Single-year VAM scores provide information about teacher effectiveness
- Changes in scores over time reflect changes in teacher effectiveness and statistical noise

EOC Model Options

AIR®
American Institutes for Research®

Objectives

- Following the SGIC's direction, we have implemented different analyses for the EOC to see if new methods can improve on the models
- Focus on the grade 9 Algebra EOC model in order to make comparisons to implemented model
- The primary aim is to determine if other models can improve on the approach previously used for the EOC

General Evaluation Criteria

- The following questions will be used to guide evaluation of the new models
  - Do the models implement a statistical approach that reasonably estimates teacher contributions to student learning?
    - The first question will be evaluated via your judgment - we will provide a model description along with benefits and risks of the different approaches
  - Do the statistical results (e.g., R square) indicate good model fit and conform to expectations?
    - To be evaluated through data summarizing the model - variance components, r-square, precision
  - Do the results of the models show differences across different classroom populations?
    - To be evaluated on the basis of impact data

Goals of a Value-Added Model (VAM)

- Goal is to control for "sorting" of students into classes
- Necessary because students are not randomly assigned into future classes
- If sorting is not controlled, teachers will have an advantage or disadvantage based on **who** they teach
  - Referred to as selection bias
- To measure teacher contributions to student learning, analysis should control for sorting to mitigate any effects associated with non-random assignment

How to Control for Sorting

- In FCAT VAM (and all VAMs), selection bias is statistically controlled for by prior student test scores and possibly other demographic measures
- Impact data from the FCAT and approved Algebra EOC models provide evidence that sorting is well controlled in those models and that effects associated with selection bias seem to be mitigated

Prior EOC Model Analysis

- Other EOC models (beyond Algebra grades 8 and 9) were not recommended because evidence suggested that sorting was not well controlled by prior test scores and other covariates
  - We observed “reversals” in the variance component patterns
  - Impact data showed very high correlations between teacher scores and classroom composition
  - R squares and precision were very low

New EOC Analysis

- To address these issues, we have experimented with new models that analyze the data in different ways
- The aim is to determine if a different modeling strategy can improve on the approach that has been used in Florida to date

New EOC Models

AIR®
American Institutes for Research®

Models 1-3: Enhanced Covariate Adjustment Model

- Some researchers have proposed that high school students are often sorted into different academic tracks
- If this tracking is correlated with sorting, then it would be necessary to control for course tracking to mitigate selection bias
- In Models 1-3, we control for students' prior math courses in addition to their prior test scores

Baseline Model (Model 1)

- This model is the same covariate adjustment model used for the grade 9 Algebra EOC
- The model includes teacher and school random effects
- It only uses prior test scores and mean prior as control variables
- No other covariates are used in this approach (though they are used in the operational model)
- Our aim is to assess if controlling for tracking improves on a model that controls for test scores alone, so this is a direct way to test that hypothesis

Controlling for Prior Course History

- In model 2, we explicitly control for tracking by using a student's prior course history as controls
- In this model, we treat prior course history as a random effect
- There are two reasons to use random effects as we do here:
  - First, some courses have very few students enrolled in them
  - Second, we do not have the entire population of courses, only a **sample** from the population of all possible courses

Controlling for Prior Course History

- In model 3 we also control for prior course history just as we do in model 2
- However, in this approach all prior courses are treated as fixed effects instead of random effects
  - The assumption here is that we do in fact have the population of prior courses
- The fixed effects approach is consistent with similar research
  - However, we create an indicator for each prior course and do not categorize the prior courses into “low”, “medium”, “high” courses

Grade 8 Math Courses of Students Taking the Algebra I EOC in Grade 9

- Algebra I (3.4%)
- Algebra I Honors (2.5%)
- Algebra Ia (5.6%)
- M/J Intensive Mathematics (MC) (11.4%)
- M/J Mathematics 3 (56.4%)
- M/J Mathematics 3, Advanced (19.3%)
- Others (1.4%)

Summary of Models 1-3

- Model 1:
  - Control for two prior test scores
  - Control for mean prior score of students in class
  - School and teacher random effects
- Model 2:
  - Model 1 + prior course random effects
- Model 3:
  - Model 1 + prior course fixed effects

R-Square Values

- *R*-squared is one indicator of model fit.
  - The closer the value is to 1, the better the model predicts outcome scores
  - Most FCAT R-squares are around 0.7
- The R-square values from the three models are:
  - Baseline: 0.49
  - Course Random Effects: 0.49
  - Course Fixed Effects: 0.50

Evaluating Precision

- Another measure of a model is its *precision*, or the degree to which it is able to distinguish among teachers on the basis of effectiveness
  - Values closer to zero indicate more precision
- The pseudo-reliability statistics for the three models are:
  - Baseline: 0.43
  - Random Effects: 0.44
  - Fixed Effects: 0.43

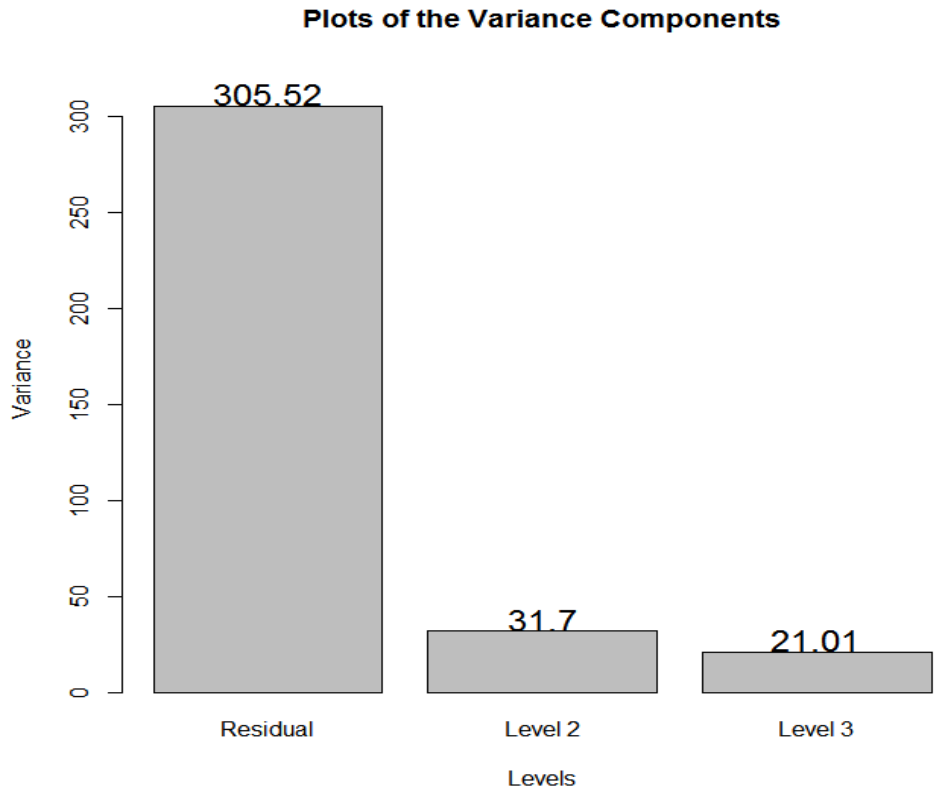
Coefficients on Model Covariates

	Model 1	Model 2	Model 3
(Intercept)	86.27	94.32	95.35
Grade 7 Prior	0.22	0.17	0.17
Grade 8 Prior	1.16	1.18	1.18
Missing Gr. 7	50.63	39.56	39.17
Mean Prior	-0.06	-0.06	1.24

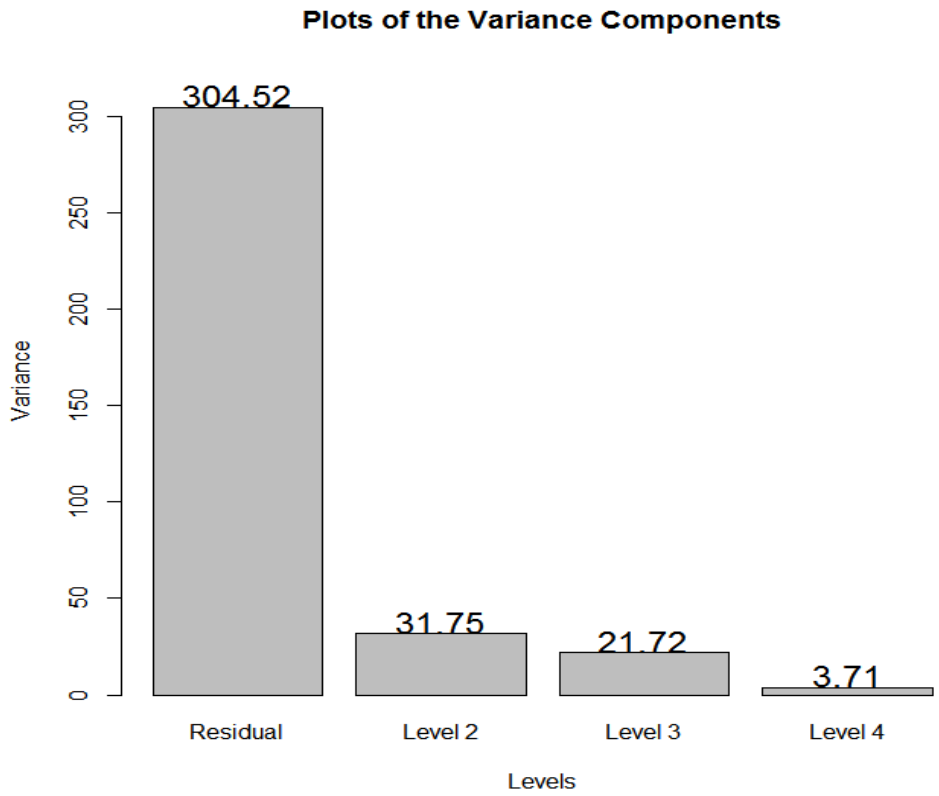
Variance Components

- Variance components summarize the existing differences between students, teachers, and schools
- If prior course history has an impact we might observe:
  - Residual variance will decrease relative to the baseline model
  - Teacher variance component might increase relative to the baseline model
  - Variance between students in different courses might explain a large share of the overall variation

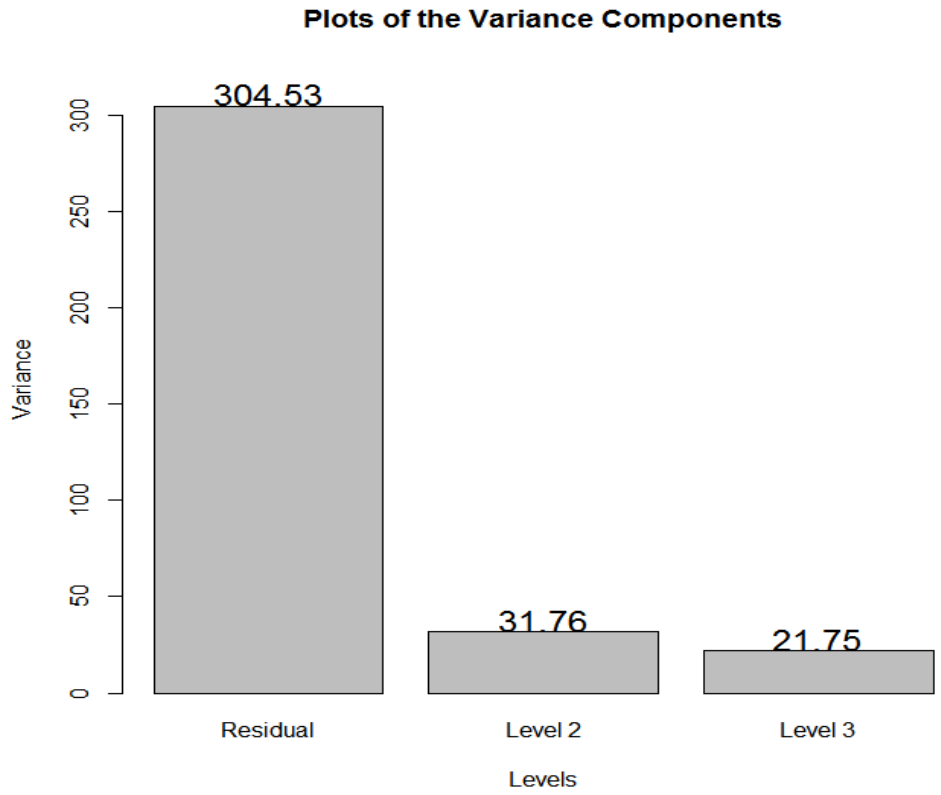
Baseline Model Variance Components



Enhanced Model Variance Components (Random Effects)



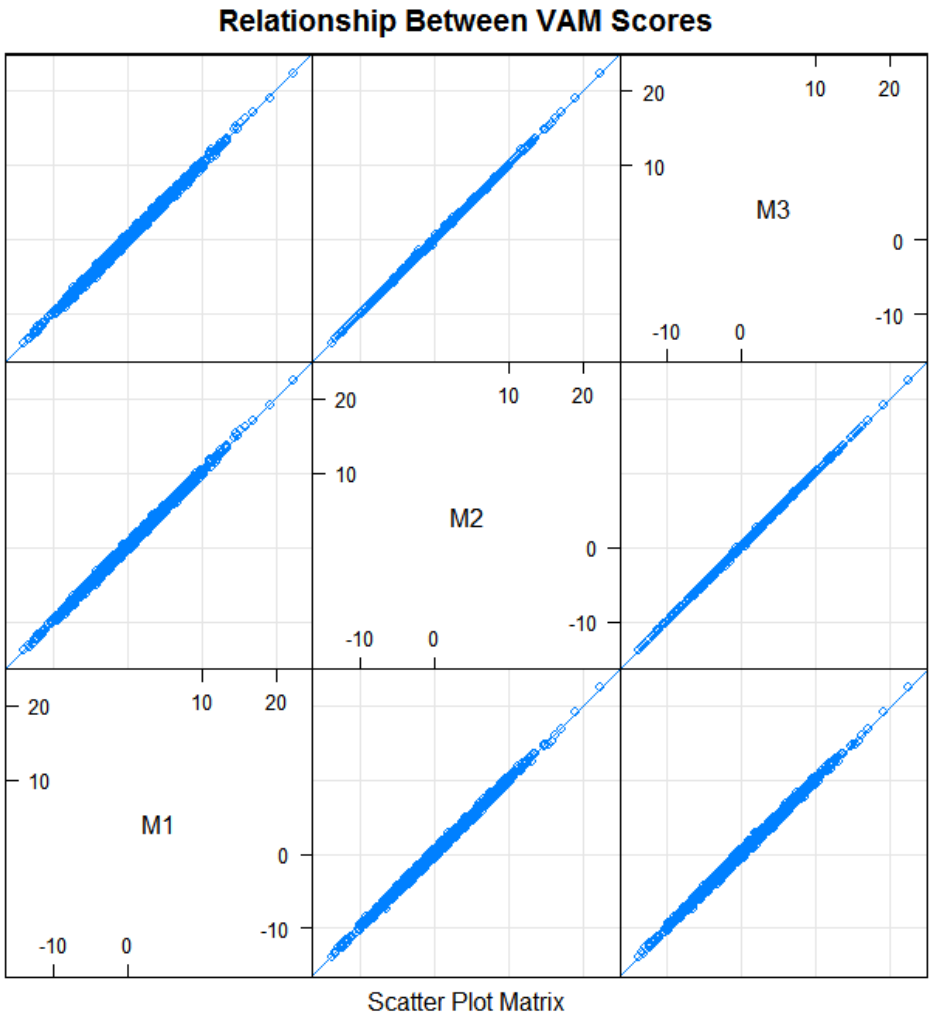
Enhanced Model Variance Components (Fixed Effects)



Summary of Variance Components

- The residual variance is virtually the same in all three models (noise is not reduced)
- The teacher variance component is also stable over all three approaches
- Variance between students taking different courses is extremely small (3.71)

Correlation Between Teacher VAM Scores



Impact Correlations

	Approved Grade 9 Algebra Model	Baseline	Random Effects	Fixed Effects	Z-Score	Pct. Prof	Prob. Prof
Mean Prior	0.058	0.093	0.095	0.095			
% Low Income	-0.043	-0.087	-0.086	-0.086			
% SWD	-0.035	-0.081	-0.083	-0.082			
% ELL	0.041	0.094	0.091	0.090			
% Non-White	-0.003	-0.007	-0.009	-0.009			

Overall Summary of Course History Models

- Statistical models yield identical results
  - R-squares and precision values are equal
  - Variance components do not change
  - Correlation in teacher VAM scores is 0.99
  - Model coefficients are comparable except for mean prior
- Impact statistics in models 2 and 3 are no different than the baseline approach

Similar Models Were Implemented for the Geometry EOC

- Models were implemented separately for grade 9 and grade 10
- Three models were run for each grade
  - The baseline model including only prior scores as covariates
  - A model that includes course histories as random effects
  - A model that includes course histories as fixed effects
- The conclusions form these models were the same as for the Algebra I EOC: controlling for course history adds almost no explanatory power to the models

Overall Summary of Course History Models

- Adding course history to the model adds little explanatory power
- True in most VAM models; once we control for prior test scores, additional covariates provide little new information
- Inability to control effectively for sorting in the EOC models is not a shortcoming of the models themselves, but with the data
- Neither the correlations between the prior test scores and the EOC outcome scores nor the correlations between the prior test scores and the curricular content of EOC courses are large enough to control effectively for sorting

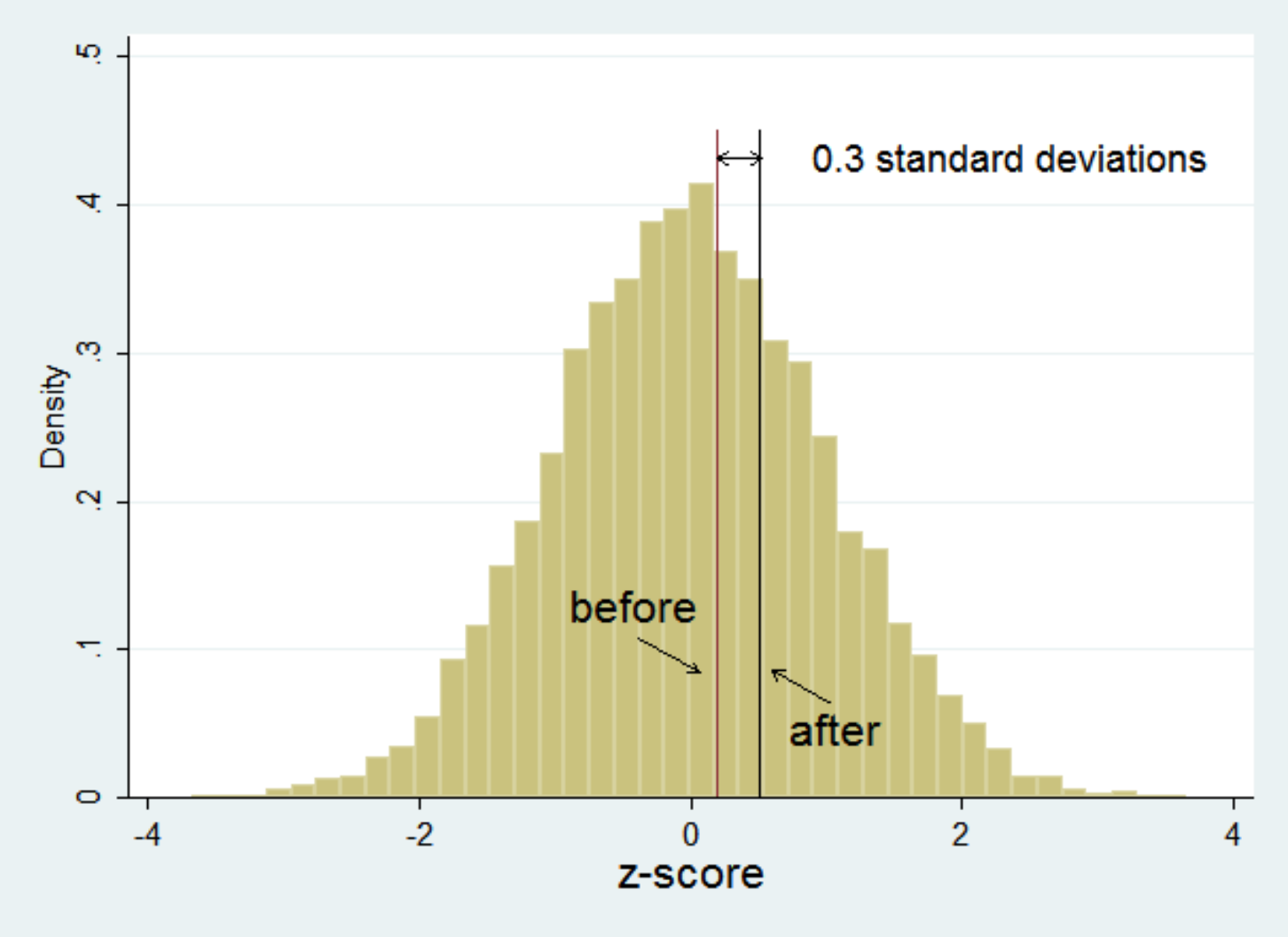
Introduction to Models 4-6

- These models are not covariate-adjustment models
  - Therefore the statistical summaries previously presented do not apply
- Model 4: Z-Score Model
  - How much do the teacher's students move up/down relative to other students?
- Model 5: Percent Achieving Proficiency
- Model 6: Probability of Proficiency
  - Measures impact of teacher on the probability the student achieves proficiency on Algebra I EOC

Model 4: Z-Score Model

- Measure where in the overall distribution of student scores each student's grade 8 math score falls
- Measure where in the overall distribution of student scores each student's Algebra I EOC score falls
- Compare the two for each student to determine how much the student moved up or down in the overall distribution of student scores
  - Positive: moved up in the distribution
  - Negative: moved down in the distribution
  - Zero: stayed in the same place relative to other students

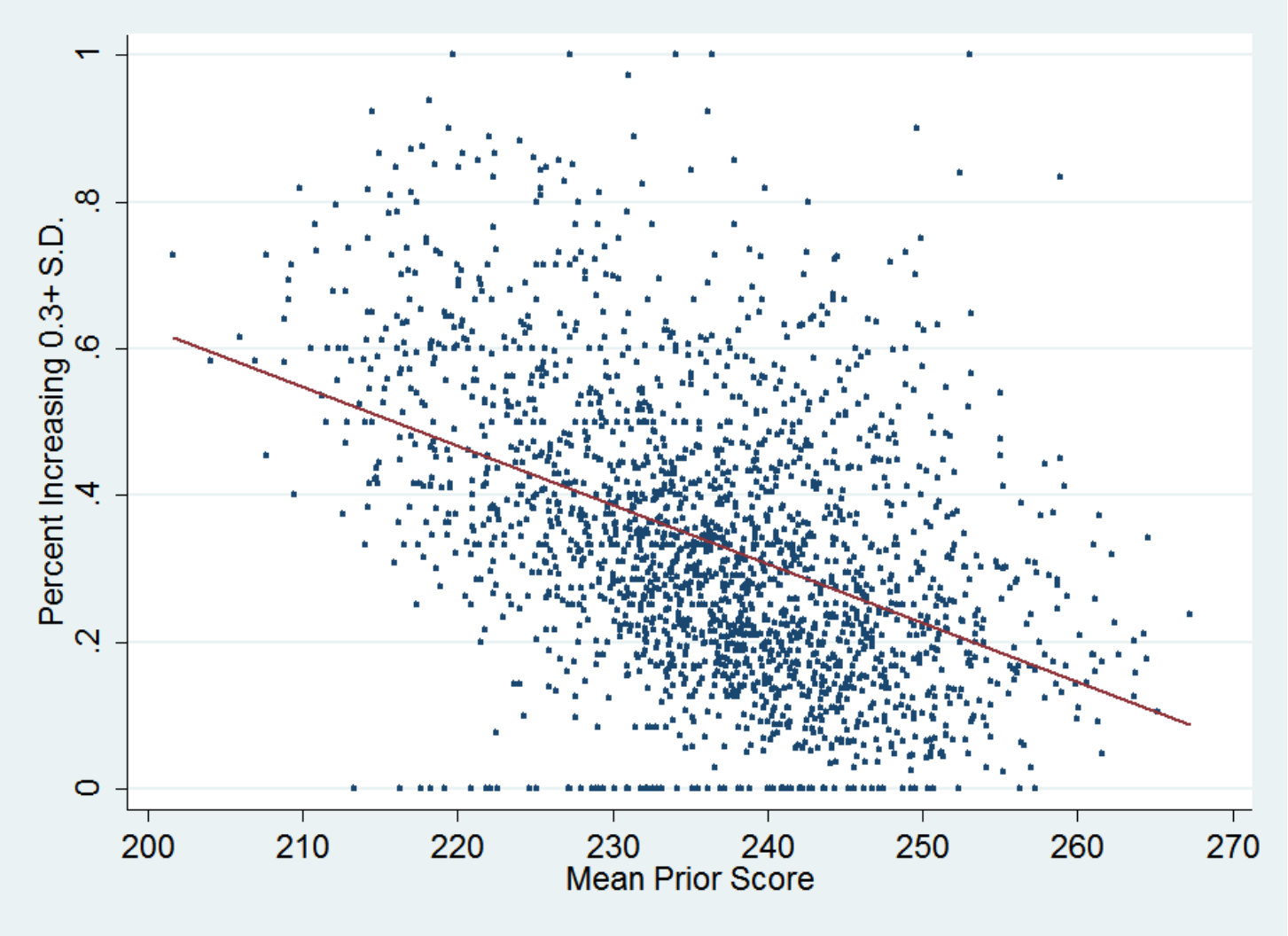
Student EOC Scores Converted to Z-Scores



Model 4: Z-Score Model

- Teacher's score = share of students who move up more than 0.3 standard deviations (s.d) in the distribution
- Assumes all students are equally likely to move up 0.3 s.d. conditional on their prior scores.
- Relatively difficult for students with high grade 8 scores to move up 0.3 s.d.
- Relatively easy for students with very low grade 8 scores to move up 0.3 s.d. (due to measurement error)
- Unlike Model 4 (percent achieving proficiency), Model 5 puts teachers of students with high grade 8 scores at a disadvantage

Model 4: Z-Score Model



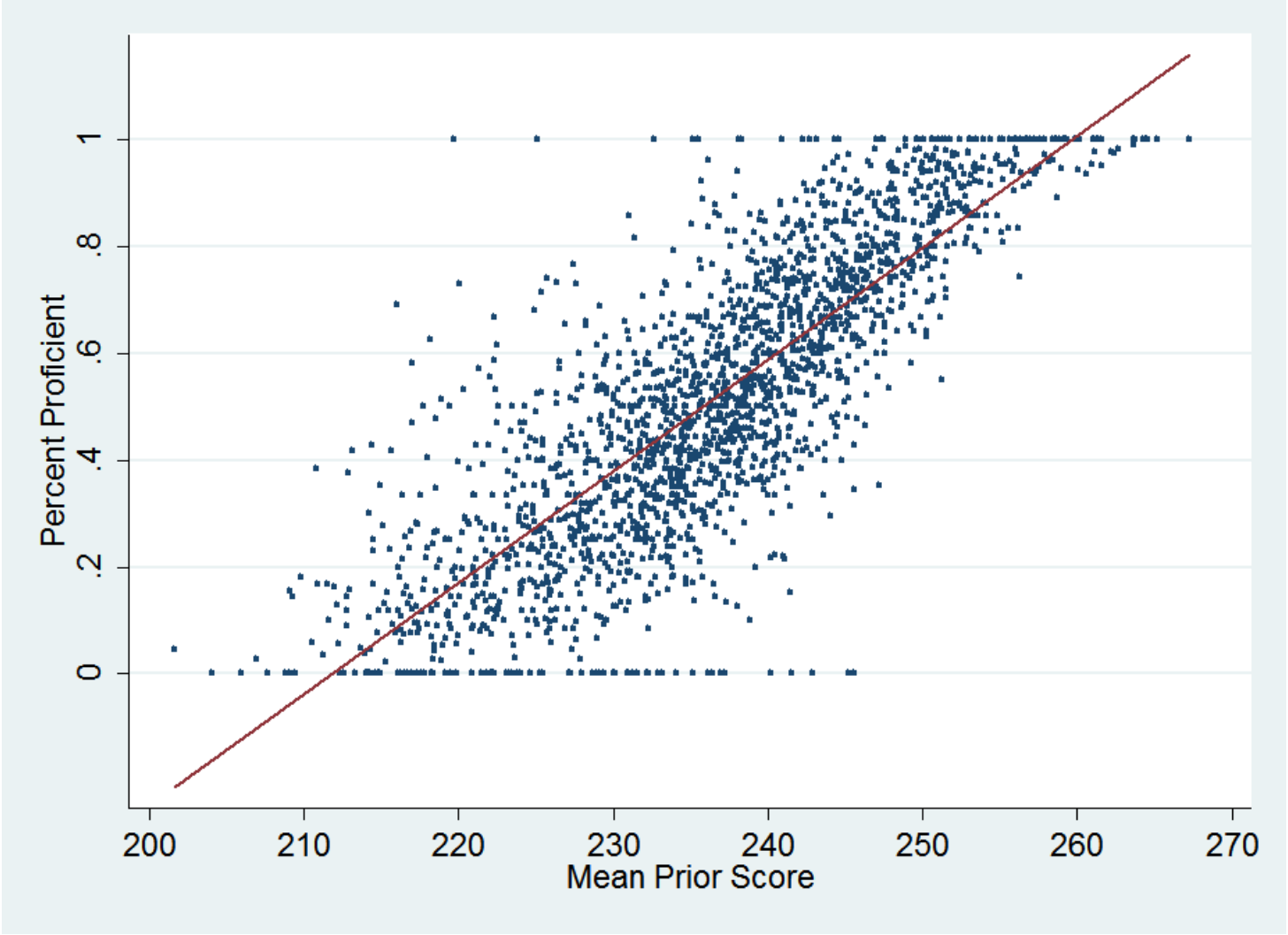
Impact Correlations

	Approved Grade 9 Algebra Model	Baseline	Random Effects	Fixed Effects	Z-Score	Pct. Prof	Prob. Prof
Mean Prior	0.058	0.093	0.095	0.095	-0.402		
% Low Income	-0.043	-0.087	-0.086	-0.086	0.173		
% SWD	-0.035	-0.081	-0.083	-0.082	0.176		
% ELL	0.041	0.094	0.091	0.090	0.198		
% Non-White	-0.003	-0.007	-0.009	-0.009	0.160		

Model 5: Percent Achieving Proficiency

- Approach commonly associated with AYP
- Teacher rating is the share of students achieving proficiency (scoring above 399)
- Does not control for sorting
- Assumes students are randomly distributed across schools
- Does not control for prior test scores or any other covariates

Model 5: Percent Achieving Proficiency



Model 5: Percent Achieving Proficiency

- Teacher scores are highly correlated with students' prior scores
- Models such as this are useful in accountability systems when the emphasis is primarily based on identification of classrooms where students achieve a passing score
- These models typically provide different information about classrooms than is observed with growth models, but the percentage of students achieving proficiency is still a valuable outcome

Impact Correlations

	Approved Grade 9 Algebra Model	Baseline	Random Effects	Fixed Effects	Z-Score	Pct. Prof	Prob. Prof
Mean Prior	0.058	0.093	0.095	0.095	-0.402	0.807	
% Low Income	-0.043	-0.087	-0.086	-0.086	0.173	-0.378	
% SWD	-0.035	-0.081	-0.083	-0.082	0.176	-0.456	
% ELL	0.041	0.094	0.091	0.090	0.198	-0.145	
% Non-White	-0.003	-0.007	-0.009	-0.009	0.160	-0.244	

Model 6: Probability of Proficiency

- Use a student's prior test scores to estimate the probability the student will score above the proficiency cut-point
- Students with higher prior test scores have a higher predicted probability of passing
- Other covariates (SWD status, ELL status, prior course history, etc.) can be included in the model as well
- Conditional on a student's prior test scores (and possibly other covariates), we can determine whether some teachers' students are more likely to pass than other teachers' students

Alternative Probability of Prediction Methods Possible

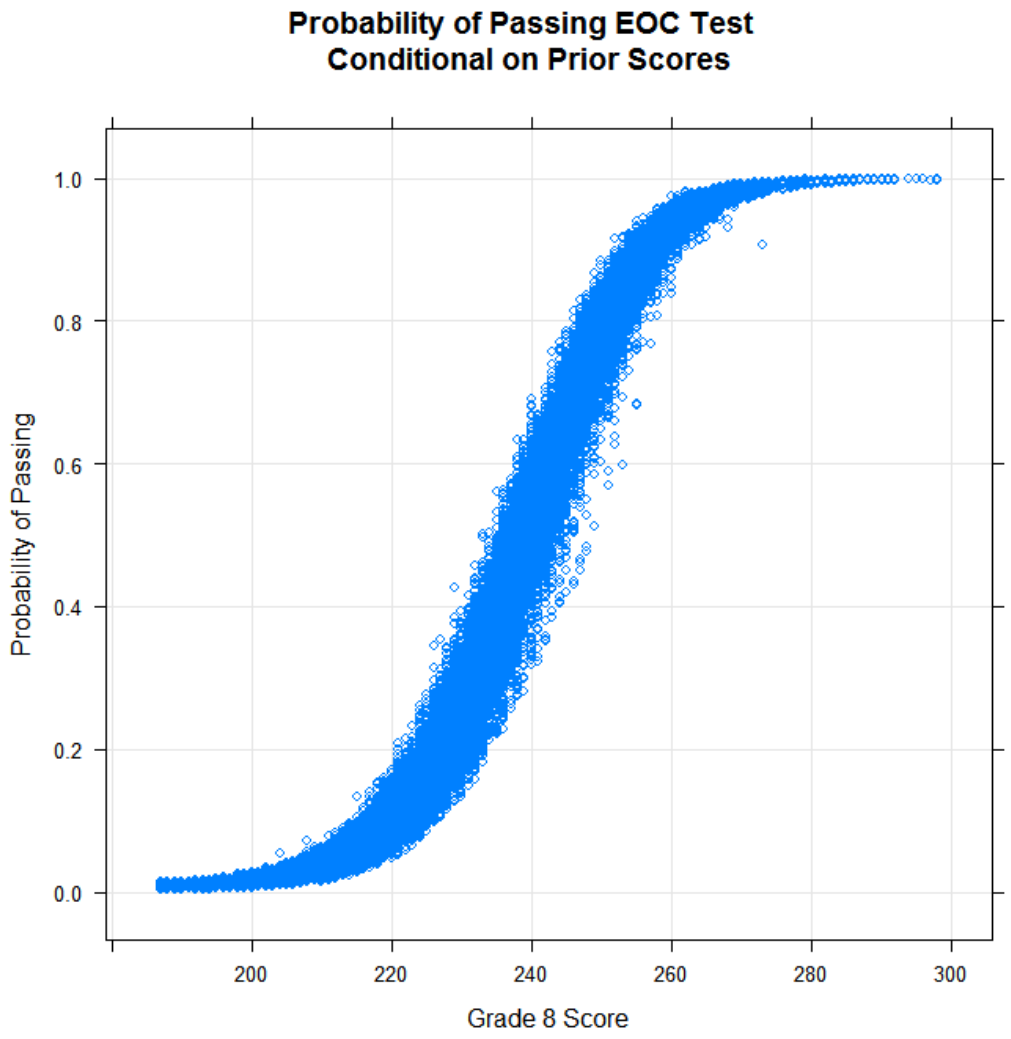
Prior FCAT Math Level	Prediction
	% Who Pass Algebra EOC
Level 1	10%
Level 2	25%
Level 3	60%
Level 4	85%
Level 5	95%

As an example, predictions could be based on percentage of students who pass based on statewide data or district data

Model 6: Probability of Proficiency

- Model assumes that conditional on prior test scores and other included covariates, students are randomly distributed across teachers and schools
- If on average a teacher's students had a low probability of passing, but many of these students passed the cut-off, that teacher would receive a high score
- If a teacher's students pass or do not pass about as expected, that teacher would receive an average score
- If fewer of a teacher's students passed than was expected, based on their prior test scores, that teacher would receive a low score

Probability of Proficiency Model



Compare Actual to Predicted

- Share of outcomes correctly predicted is one measure of model fit
- Model correctly predicts passage for 77% of students

Pass Rates		Actual	
		Not Pass	Pass
Predicted	Not Pass	34678 (36.1%)	12620 (13.1%)
	Pass	9117 (9.5%)	39612 (41.3%)

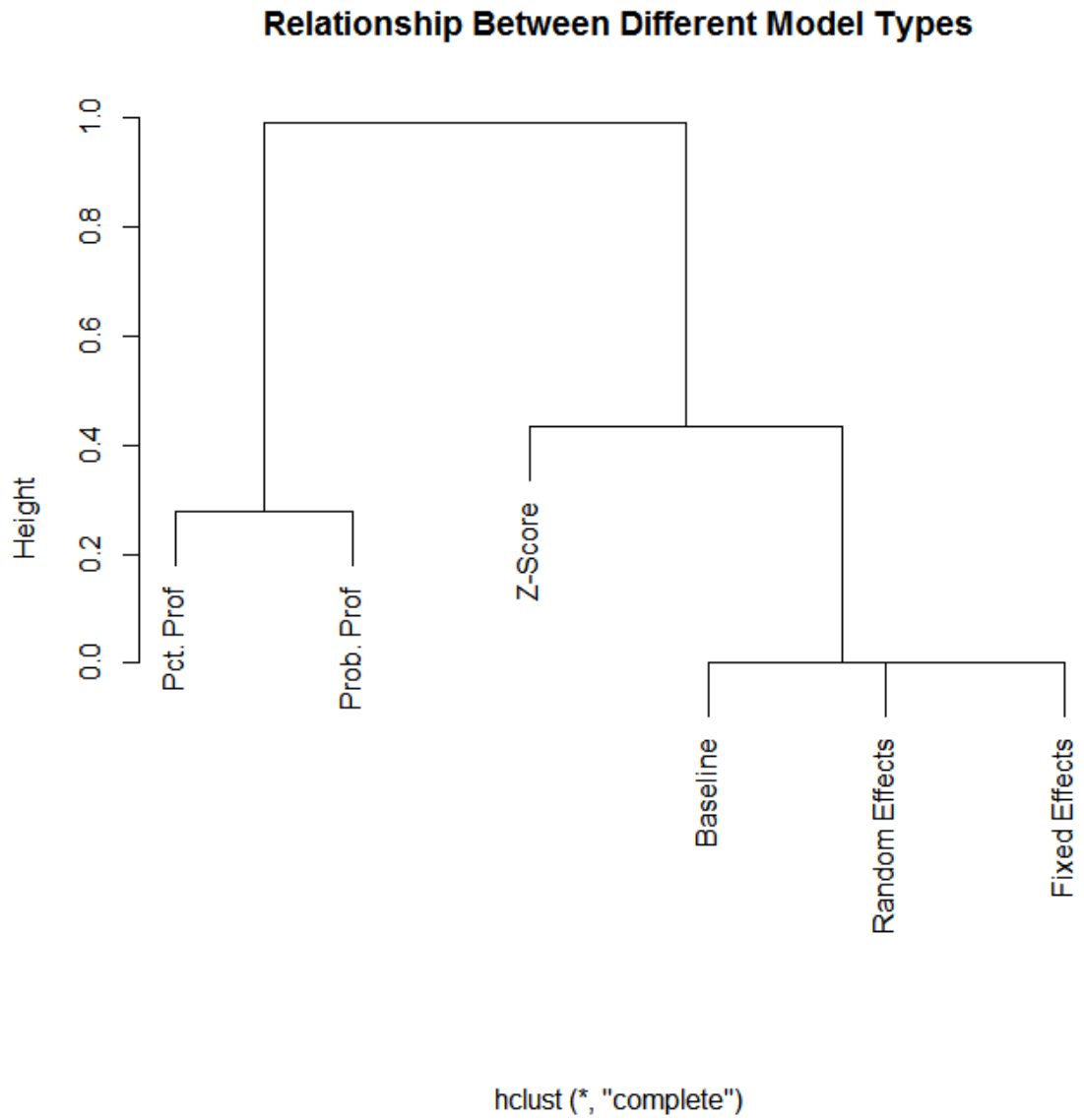
Impact Correlations

	Approved Grade 9 Algebra Model	Baseline	Random Effects	Fixed Effects	Z-Score	Pct. Prof	Prob. Prof
Mean Prior	0.058	0.093	0.095	0.095	-0.402	0.807	0.243
% Low Income	-0.043	-0.087	-0.086	-0.086	0.173	-0.378	-0.127
% SWD	-0.035	-0.081	-0.083	-0.082	0.176	-0.456	-0.193
% ELL	0.041	0.094	0.091	0.090	0.198	-0.145	0.058
% Non-White	-0.003	-0.007	-0.009	-0.009	0.160	-0.244	-0.050

Summary of Models 4-6

- Model 4 (Z-Score):
  - Rewards teachers whose students make significant growth in the overall distribution of student scores
  - Disadvantages teachers whose students have *high* math 8 scores
- Model 5 (Percent Achieving Proficiency):
  - Measures share of students who achieve proficiency
  - Similar to AYP
  - Disadvantages teachers whose students have low math 8 scores
- Model 6 (Probability of Proficiency):
  - Measures teachers' impact on the probability a student achieves proficiency
  - Has advantages similar to covariate adjustment model

Similarity Graph (Correlations) Between Models



Correlations Between Models

	Baseline	Random Effects	Fixed Effects	Z-Score	Pct. Prof	Prob. Prof
Baseline	1	0.999	0.999	0.569	0.423	0.618
Random Effects	0.999	1	0.99	0.567	0.424	0.616
Fixed Effects	0.999	0.99	1	0.567	0.424	0.616
Pct. Prof	0.423	0.424	0.424	0.007	1	0.721
Z-Score	0.569	0.567	0.567	1	0.007	0.489
Prob. Prof	0.618	0.616	0.616	0.489	0.721	1

Summary

- Controlling for students' prior courses does little to improve predictive power of covariate adjustment models
- "Percent achieving proficiency" and z-score models do not control for sorting
- The benefits of the "probability of proficiency" models come close to those of the covariate adjustment models

Harold Doran and Eric Larsen
hdoran@air.org and slarsen@air.org

1000 Thomas Jefferson Street NW
Washington, DC 20007
General Information: 202-403-5000
TTY: 887-334-3499
www.air.org

AIR®
American Institutes for Research®