RAND EDUCATION

Are Teachers Differentially Effective with Students of Differing Abilities?

J.R. Lockwood & Daniel F. McCaffrey

April 24, 2008

This material is based on work supported by the Department of Education Institute of Education Sciences under Grant No. R305U040005 . Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of these organizations.

Goal of Value-Added Modeling (VAM) is to Provide Fair Comparisons of Teachers Teaching Different Kinds of Students

- **Premise is that rich longitudinal data on students, combined with sophisticated modeling, can be used to approximate a randomized experiment**
  - **Groups of students taught by different teachers differ with respect to backgrounds and prior achievement**
  - **Hope is that statistical adjustment can remove these differences**
- **Much of VAM literature has focused on how well different models can provide estimates of individual teacher effects that are free of bias from student characteristics**
  - **I.e. internal validity of teacher effect estimates**

But How Generalizable Are Teacher Effect Estimates?

- Assume internal validity
- Most models estimate a single effect for each teacher, perhaps separately by subject and by year when data permit
- To what units (students), outcomes (subjects and tests), and settings (schools, years, courses, other contextual factors) do VAM estimates provide a generalizable inference?

This Work Examines Heterogeneity of Teacher Effects Across Different Students

- **Anecdotal evidence and experience suggests that teachers might be differentially effective with different types of students**
  - **E.g. different aptitudes or other characteristics**
- **We focus on whether teacher effects vary as a function of a student's "general ability level"**
  - **Latent characteristic that can be inferred from extensive longitudinal achievement data series providing measures of an individual student's achievement taken from different grades, subjects and contexts**
- **Related to prior work (e.g. Sanders, Koedel and Betts, Hanushek et al.) finding mixed results on student-teacher interactions**

Understanding Student-Teacher Interactions is Important to VAM

- **Challenge: Heterogeneity of a teacher's effects across students erodes generalizability of VAM estimates**
  - **Models with a single effect are implicitly estimating the effect of each teacher *on his/her students***
  - **Two teachers who would be equally effective on similar students could get systematically different estimates**
- **Opportunity: If heterogeneity can be reliably measured, could enhance utility of VAM for improving education**
  - **Diagnostic information for targeted interventions**
  - **More efficient assignments of student/teacher pairings to take advantage of teachers' individual strengths**

Outline

- Develop model that allows individual teacher effects to vary as a function of students' general ability levels
- Use variety of data sources to examine:
  - Is there evidence for student-teacher interactions?
  - How big are they?
  - What are the implications for ignoring them?

A Basic Model for Student-Teacher Interactions

$$Y_i = \mu + \delta_i + \theta_{0j(i)} + \theta_{1j(i)}\delta_i + \epsilon_i$$

- **Simplified scenario:** $i = 1, \ldots, I$ **students each uniquely linked to one of** $j = 1, \ldots, J$ **teachers (think one school year)**
- $j(i)$ **denotes the teacher** j **who teaches student** i
- δ_i **- "general ability" of student** i**, scaled so that it has mean zero across all students but otherwise is on the scale of the test**
- θ_{0j} **- the main effect of teacher** j
  - **Mean zero across teachers**
- θ_{1j} **- interaction effect of teacher** j
  - **Mean zero across teachers**
  - $\theta_{1j} > 0$ **indicates teacher** j **is relatively more effective with students of above-average ability**

Use Prior Test Scores to Estimate Student Abilities

- **Suppose p prior scores on students, from previous grades and potentially coming from different subjects**

$$Z_{ip} = \mu_p + \beta_p \delta_i + \epsilon_{ip}$$

- **Each prior score depends on δ_i but perhaps differentially through β_p**
- **Residuals ϵ_{ip} assumed independent with variance depending on p**
  - **Probably misspecified because of ignoring prior teacher effects, subject-specific abilities, and persistent heterogeneity in growth**
  - **Paper presents an auxiliary analysis that suggests that misspecification is not affecting substantive conclusions**
- **Basically providing a generalized average prior score, put onto the scale of the target score Y_i**

Joint Model for Target Year Score and Prior Scores Estimated in Bayesian Framework

$$Y_i = \mu + \delta_i + \theta_{0j(i)} + \theta_{1j(i)}\delta_i + \epsilon_i \quad (1)$$

$$Z_{ip} = \mu_p + \beta_p\delta_i + \epsilon_{ip} \quad (2)$$

- **Estimated in WinBUGS**
- **Accommodates arbitrary missingness patterns of prior scores via MAR assumption**
- **Key variance components estimated from the data:**
  - **Var**(δ), **Var**(θ_0), **Var**(θ_1), **Cor**(θ_0, θ_1)
- **In (1) we extended model to handle nonlinearity of mean structure, and heteroskedasticity of** ϵ_i

Empirical Analyses Based on Data from Three Large Urban School Districts

- Examine interactions for twelve “target teacher groups”
  - Six from District A: math and reading from grades 3, 4, 5
  - Two from District B: math from grades 7 and 8
  - Four from District C: math from grades 5, 6, 7, 8
- Between 35 and 320 teachers in each target teacher group
- Between 1400 and 4500 students linked to target teachers
- Between 4 and 14 prior scores available for estimating δ_i
- District B is pivot between Districts A and C: math test outcome is the same as District A but grade ranges overlap with District C

Results: Model Selection Criteria Favor the Model Allowing for Student-Teacher Interactions

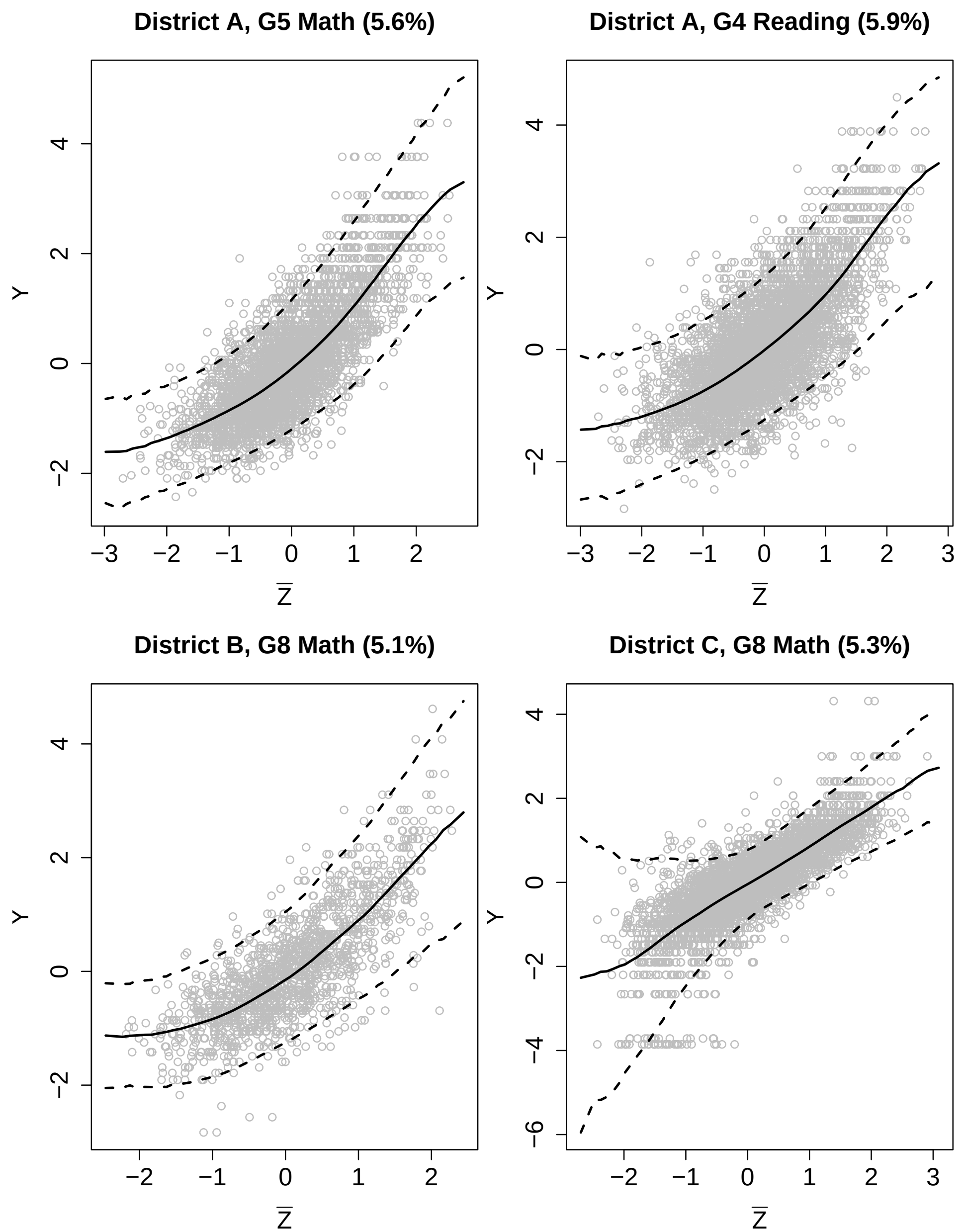
- Interaction terms improve the performance of the model relative to simpler alternatives for all twelve target teacher groups

- Compared sequence of five increasingly complex models using Deviance Information Criterion (DIC):

	Nonlinearity	Heteroskedasticity	Tch Main Effx	Interactions
1	No	No	No	No
2	Yes	No	No	No
3	Yes	Yes	No	No
4	Yes	Yes	Yes	No
5	Yes	Yes	Yes	Yes

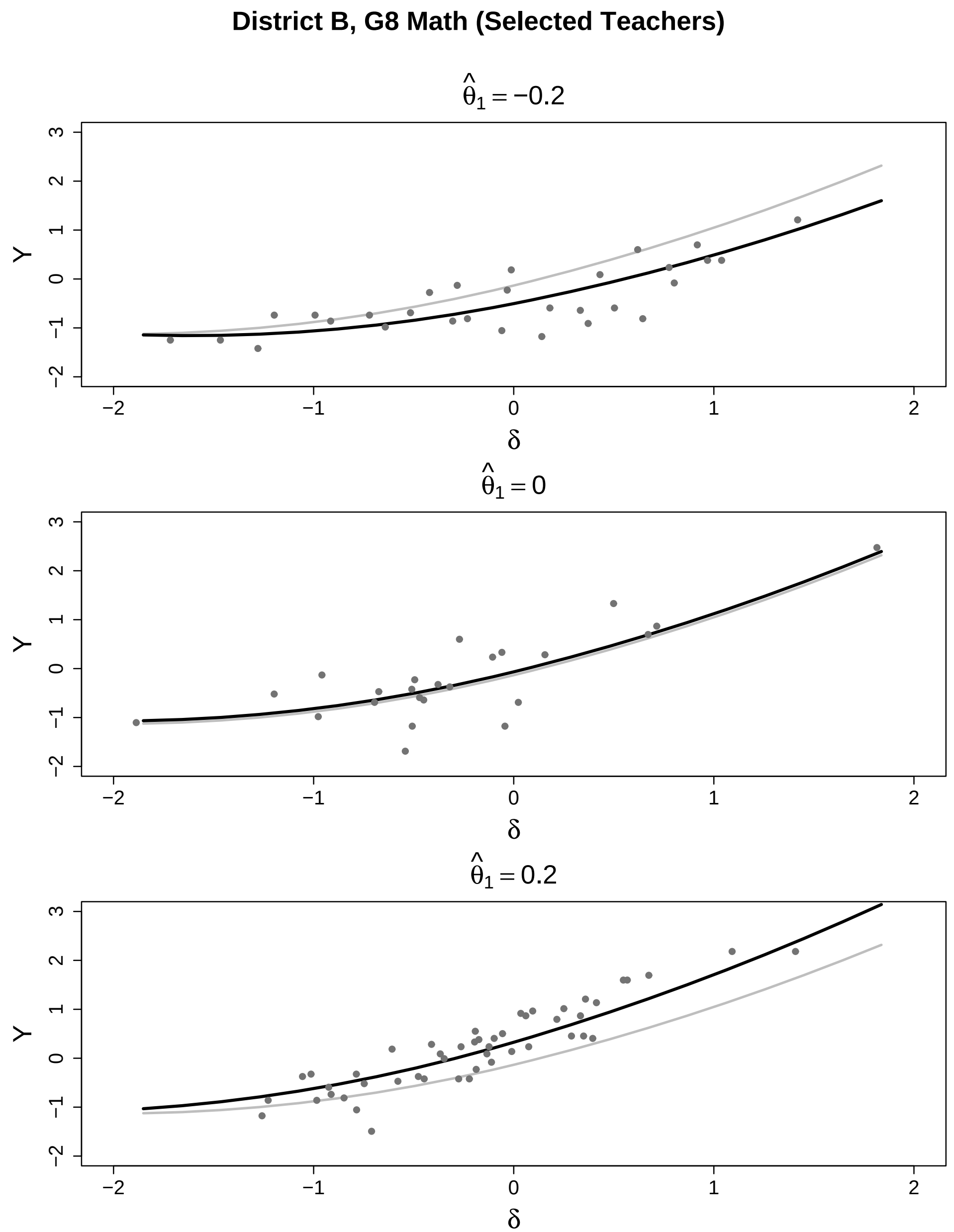
- Model 5 (“complete model”) preferred in all cases

Complete Model Fits the Data Well



RAND

Complete Model Fits the Data Well



How Large Are Teacher Main Effects?

- Consistent with previous research
- Var(θ_0) on the order of about 10% of the marginal variance of Y_i; somewhat higher in early elementary grades of District A

How Large Are Interactions?

- **Harder to calibrate because depends on both Var(θ_1) and Var(δ)**
- **We use**

$$\gamma = 100 \times \frac{\textbf{Var}(\theta_1)\textbf{Var}(\delta)}{\textbf{Var}(\theta_0) + \textbf{Var}(\theta_1)\textbf{Var}(\delta)}$$

- **Loosely interpreted as percentage of the total variation in potential effects of teachers on individual students that is due to interactions**
- **Consistent across 10 of 12 teacher groups at around 10%**
  - **(District C grades 7 and 8: 25% and 15% respectively)**

How Much Interactions Matter in Practice Depends on How Heterogeneous Classes Are

- Under the model the average effect that a teacher has on a class is

$$\theta_0 + \theta_1 \bar{\delta}$$

- 20-30% of the variance in δ lies between teachers for Districts A and B, with notably larger values between 50 and 60% for District C
- Model estimates that only between 2%-4% of variance in estimated teacher effects is due to interactions
  - (District C grades 7 and 8: 15% and 9% respectively)

Suggests Ignoring Interactions Would Not Lead to Appreciable Differences for Most Teachers

- But might not be true in all contexts (grades 7 and 8 in District C)
- And teachers with large interaction effects, and who are assigned to classes with relatively extreme $\bar{\delta}$, could receive estimates that are substantially different than had the teachers taught a much different class

How Are Interactions Related to Main Effects?

- Do above-average teachers perform relatively better or worse with above average students?
- This correlation is estimated by the model
- Positive in Districts A and B, and either zero or negative for District C
  - Recall Districts A and B have the same test
- Rescaling the tests using rank-based z-scores changes the correlations and estimated interaction effects, but has almost no impact on main effects
- $\Rightarrow$ Inferences about interactions can be highly sensitive to the properties of the test score scale

Conclusions

- A model allowing for teacher interactions with students' general abilities appears better for the data than simpler alternatives
- Model suggests that 2%-4% of variance in average teacher effects for their given classes is due to these interactions
- Ignoring interactions is probably not misleading for most teachers in most settings, but results suggest that potential exists
- Main effects are virtually invariant to rescaling but features of interactions are not
  - Further research on these sensitivities is required (e.g. more complex measurement error models)