RAND EDUCATION

Challenges for Value-Added Assessment of Teacher Effects

Daniel F. McCaffrey

October 22, 2004

Value-Added Models of Student Achievement

- Goal of value-added models (VAM) is to estimate causal effects of individual schools and teachers on student growth
  - Assumes schools and teachers contribute to growth
  - Recognizes that level of achievement is highly dependent on non-educational inputs
  - Attempts to separate schooling from other inputs to achievement (e.g., family or student background)
- VAM uses statistical analysis of repeated scores from students to estimate teacher effects

Policy Makers are Very Interested in VAM

- **Several states provide school districts with VAM results, including TN, PA, OH, IA**
  - **Except for TN, use of VAM is within the last couple of year**
- **More states are considering implementation (LA, DE, NY and CA)**
- **Advocates have proposed using VAM results for teacher promotion, retention, and salary decisions**

High Expectations for VAM

- "[VAM] provides unbiased estimates of the effects of schooling on individual and group academic progress" (PSDE, 2002)
- "[VAM] separates annual growth [in scores] into two parts: that which can be attributed to the teacher and that which can be attributed to the student" (Hershberg, 2004)
- "[VAM] may yield answers to questions such as: Do some teachers demonstrate consistently greater/less effectiveness in adding to student assessment gains?" (PSDE, 2002)

Goals for Today's Talk

- **Investigate four challenges to estimating teacher effects using VAM**
  - **Persistence of teacher effects**
  - **Effects of student background variables**
  - **Missing scores and missing teacher links**
  - **Modeling approach**
- **Using variety of methods**
  - **Empirical analyses of data from a large urban school district**
  - **Analytic analyses**
  - **Simulation study**

Outline

- Persistence of teacher effects
- Effects of student background variables
- Missing scores and missing teacher links
- Modeling approach

A Model for Longitudinal Test Score Data (The Persistence Model)

$$y_{i1} = \mu_1 + \boldsymbol{\beta}_1' \mathbf{x}_{i1} + \boldsymbol{\lambda}_{i1}' \boldsymbol{\eta}_1 + \boldsymbol{\phi}_{i1}' \boldsymbol{\theta}_1 + \epsilon_{i1}$$

$$y_{i2} = \mu_2 + \boldsymbol{\beta}_2' \mathbf{x}_{i2} + \boldsymbol{\lambda}_{i2}' \boldsymbol{\eta}_2 + \boldsymbol{\phi}_{i2}' \boldsymbol{\theta}_2 + \omega_{21} \boldsymbol{\lambda}_{i1}' \boldsymbol{\eta}_1 + \alpha_{21} \boldsymbol{\phi}_{i1}' \boldsymbol{\theta}_1 + \epsilon_{i2}$$

$$y_{i3} = \mu_3 + \boldsymbol{\beta}_3' \mathbf{x}_{i3} + \boldsymbol{\lambda}_{i3}' \boldsymbol{\eta}_3 + \boldsymbol{\phi}_{i3}' \boldsymbol{\theta}_3 + \omega_{32} \boldsymbol{\lambda}_{i2}' \boldsymbol{\eta}_2 + \alpha_{32} \boldsymbol{\phi}_{i2}' \boldsymbol{\theta}_2 + \omega_{31} \boldsymbol{\lambda}_{i1}' \boldsymbol{\eta}_1 + \alpha_{31} \boldsymbol{\phi}_{i1}' \boldsymbol{\theta}_1 + \epsilon_{i3}$$

A Model for Longitudinal Test Score Data
(The Persistence Model)

$$y_{i1} = \mu_1 + \boldsymbol{\beta}_1' \mathbf{x}_{i1} + \boldsymbol{\lambda}_{i1}' \boldsymbol{\eta}_1 + \boldsymbol{\phi}_{i1}' \boldsymbol{\theta}_1 + \epsilon_{i1}$$

$$y_{i2} = \mu_2 + \boldsymbol{\beta}_2' \mathbf{x}_{i2} + \boldsymbol{\lambda}_{i2}' \boldsymbol{\eta}_2 + \boldsymbol{\phi}_{i2}' \boldsymbol{\theta}_2 + \omega_{21} \boldsymbol{\lambda}_{i1}' \boldsymbol{\eta}_1 + \alpha_{21} \boldsymbol{\phi}_{i1}' \boldsymbol{\theta}_1 + \epsilon_{i2}$$

$$y_{i3} = \mu_3 + \boldsymbol{\beta}_3' \mathbf{x}_{i3} + \boldsymbol{\lambda}_{i3}' \boldsymbol{\eta}_3 + \boldsymbol{\phi}_{i3}' \boldsymbol{\theta}_3 + \omega_{32} \boldsymbol{\lambda}_{i2}' \boldsymbol{\eta}_2 + \alpha_{32} \boldsymbol{\phi}_{i2}' \boldsymbol{\theta}_2 + \omega_{31} \boldsymbol{\lambda}_{i1}' \boldsymbol{\eta}_1 + \alpha_{31} \boldsymbol{\phi}_{i1}' \boldsymbol{\theta}_1 + \epsilon_{i3}$$

A Model for Longitudinal Test Score Data (The Persistence Model)

$$y_{i1} = \mu_1 + \boldsymbol{\beta}_1'\mathbf{x}_{i1} + \boldsymbol{\lambda}_{i1}'\boldsymbol{\eta}_1 + \boldsymbol{\phi}_{i1}'\boldsymbol{\theta}_1 + \epsilon_{i1}$$

$$y_{i2} = \mu_2 + \boldsymbol{\beta}_2'\mathbf{x}_{i2} + \boldsymbol{\lambda}_{i2}'\boldsymbol{\eta}_2 + \boldsymbol{\phi}_{i2}'\boldsymbol{\theta}_2 + \omega_{21}\boldsymbol{\lambda}_{i1}'\boldsymbol{\eta}_1 + \alpha_{21}\boldsymbol{\phi}_{i1}'\boldsymbol{\theta}_1 + \epsilon_{i2}$$

$$y_{i3} = \mu_3 + \boldsymbol{\beta}_3'\mathbf{x}_{i3} + \boldsymbol{\lambda}_{i3}'\boldsymbol{\eta}_3 + \boldsymbol{\phi}_{i3}'\boldsymbol{\theta}_3 + \omega_{32}\boldsymbol{\lambda}_{i2}'\boldsymbol{\eta}_2 + \alpha_{32}\boldsymbol{\phi}_{i2}'\boldsymbol{\theta}_2 + \omega_{31}\boldsymbol{\lambda}_{i1}'\boldsymbol{\eta}_1 + \alpha_{31}\boldsymbol{\phi}_{i1}'\boldsymbol{\theta}_1 + \epsilon_{i3}$$

Other Models Are Special Cases of This Model

- **The layered model (TVAAS)**
  - **No school effects**
  - **No covariates**
  - $\alpha_{tt'} \equiv 1$, **for all** $t > t'$
- **Cross-classified models**
  - **Residual error terms specified by student random growth curves (e.g., random intercepts and slopes)**
  - $\alpha_{tt'} \equiv 1$, **for all** $t > t'$

How Do Teacher Effects Persist Over Time?

- Common models assume teacher effects persist undiminished into future years of testing
  - Layered model, cross-classified models
- Our previous explorations found persistence parameters to be significantly less than 1
  - α ranged from 0.2 to 0.3 after one year and 0.1 after two years
  - Very small sample of four schools and 678 students

Fitting the Persistence Model Is Computationally Challenging

- **Covariance matrix of the vector of scores is large and sparse**
  - **Covariance matrix is not block-diagonal**
  - **Cannot be inverted using simple formulas**
- **Hierarchical model packages will fit some cross-classified models but not with persistence**
  - **Challenging to specify even the layered model in MLWin**
  - **Impossible in SAS**
  - **Recent updates to R and HLM will fit cross-classified models but not with persistence parameters**

Bayesian Approach Simplifies Computation

- Does not require inverting the complete covariance matrix
- Markov Chain Monte Carlo (MCMC) algorithm draws from posterior by taking sequential samples from conditional distributions
  - Conditional distributions become simple regression-like problems

Implementing the Bayesian Approach

- WinBUGS
  - Freeware available at
    `http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml`
  - Sufficiently fast for moderate problems
  - Too slow for large data sets
- Developed our own MCMC algorithm and implemented it in C
  - Acceptably fast for even large data sets

Application to a Large Urban School District

- One of the nation's largest urban school districts
  - Enrollment of about 75,000 students
  - Estimating teacher effects for about 1,500 teachers
- Spring testing for 1998 to 2002
- Math and reading scores
- Students in grades 1 to 5 for the tested years
- Links students across years, to teachers and to schools
- Focus on grade 1 teachers in 1998 to grade 5 teachers in 2002
- 9,295 students in the analysis sample

Persistence Parameters Are Significantly Less Than One

	Math		Reading	
	Mean	Std. Dev.	Mean	Std. Dev.
α_{21}	0.11	0.02	0.15	0.02
α_{31}	0.13	0.02	0.12	0.02
α_{32}	0.14	0.02	0.20	0.03
α_{41}	0.08	0.02	0.11	0.02
α_{42}	0.07	0.02	0.13	0.03
α_{43}	0.09	0.02	0.14	0.03
α_{51}	0.08	0.02	0.10	0.02
α_{52}	0.09	0.02	0.16	0.03
α_{53}	0.08	0.02	0.13	0.03
α_{54}	0.20	0.03	0.23	0.03

Layered and Persistence Models Yield Different Inferences for Some Teachers

- Correlation between the estimated teacher effects ranges across grades from .73 to .87 for math and .77 to .84 for reading
- Teacher variance components are much larger for the persistence model (1.3 to 6.8 times larger)
- Persistence model yields more teachers whose 90% posterior intervals exclude zero, especially for grades 2 and 3
  - Example: grade 3 math,
    - For persistence model, 105 of the 306 teachers significantly above average teacher
    - For the layered model only 57 classified as above average
    - 50 overlap

Outline

- Persistence of teacher effects
- **Effects of student background variables**
- Missing scores and missing teacher links
- Modeling approach

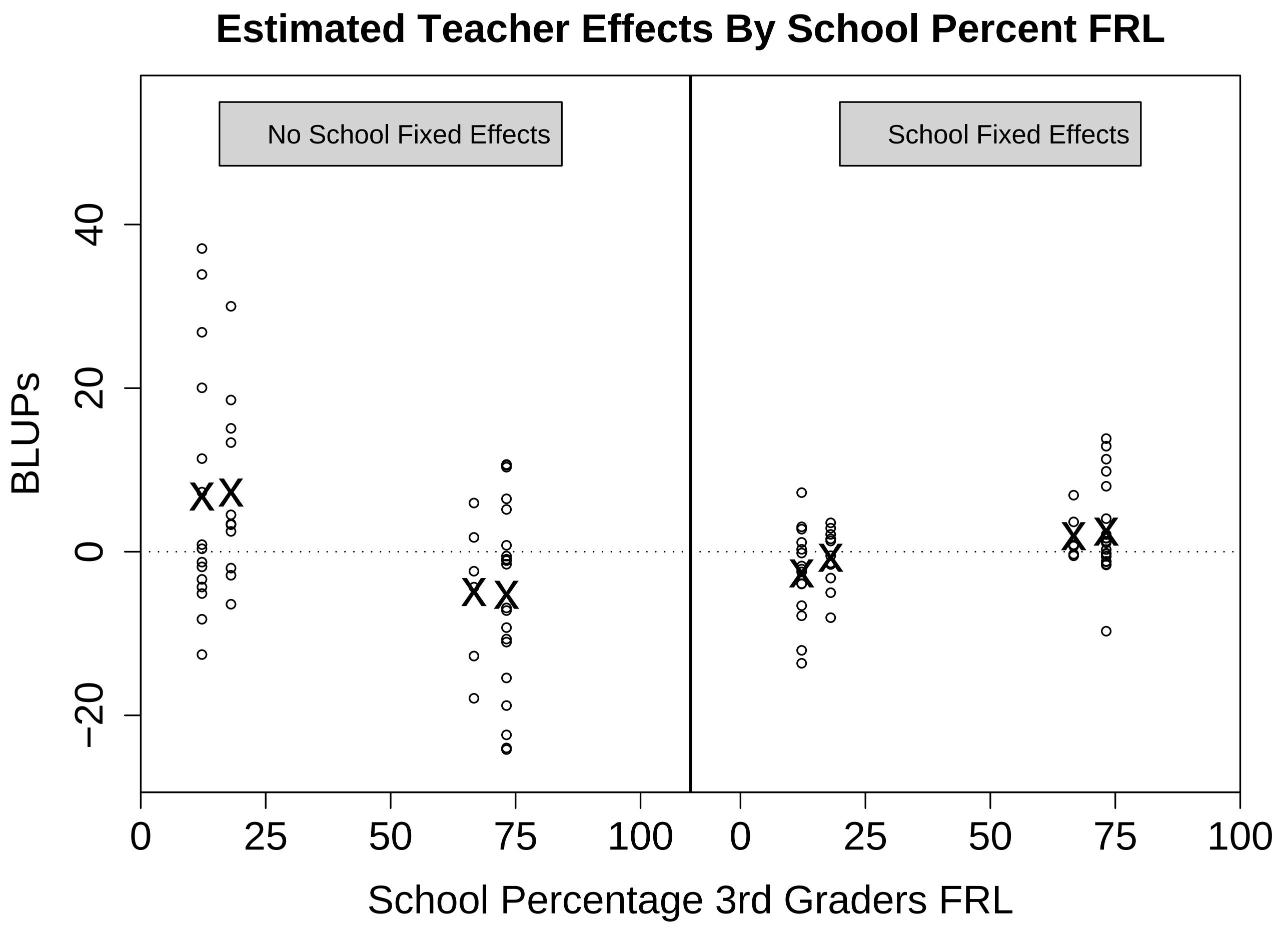
Do Estimated Effects Depend on the Students?

- Purported value of VAM is it separates teacher effects from student effects
- On average, are some estimated teacher effects higher or lower because of the students they teach?
- Do estimated effects correlate with student characteristics?

Estimated Effect Will Depend on Students But Intra-Student Correlation Mitigates Some Errors

- **Students are not uniformly grouped**
- **Some of the inter-classroom heterogeneity is associated with estimated teacher effects**
- **Accounting for intra-student correlation in the multiple scores from a student removes much of inter-classroom heterogeneity if students mix across years**
- **Accounting for intra-student correlation does *not* mitigate confounding when student populations are stratified according to achievement**

Example Where Estimated Teacher Effects Depend on the School



School Level Effects Remain an Issue in Urban School Data

- District's Title I and non-Title I schools form essentially disjoint strata of students from different populations
- Estimated teacher effects are lower on average for teachers teaching in Title I schools
  - Roughly 1/2 standard deviation unit lower in Title I Schools
- Teachers in Title I schools are much more likely to be considered below average (11 to 25% vs. 0 to 8%)
- Teachers in Title I schools are less likely to be considered above average (17 to 22% vs. 23 to 58%)

Including Student Covariates in the Model Is Not Necessarily Going to Work

- Available covariates might be incomplete
- Teacher effects are based on residuals
- Model attributes all factors correlated with covariates to covariates
  - Effects of covariate
  - Teacher effectiveness correlated with covariates
- Model over adjusts
- Estimated effects remain confounded by student characteristics

Partial Fixes for Covariate Modeling Exist

- Using within classroom variance on covariates can adjust for student level covariates without over adjusting

- Does not work for school or classroom level covariates
  - Neighborhood effects
  - Contextual effects

Outline

- Persistence of teacher effects
- Effects of student background variables
- **Missing scores and missing teacher links**
- Modeling approach

How Sensitive Are Estimated Effects to Assumptions About Missing Data?

- Missing data are pervasive
- In urban data set only 20% of observations are complete
  - 6,417 first graders in 1998
  - 370 repeated a grade
  - 29 skipped a grade
  - 2,615 transferred out of the district
  - 317 missed testing
  - 3,915 students enter the cohort after grade 1
- Observations with missing scores are also missing teacher links
- Must make assumptions about missing links and scores

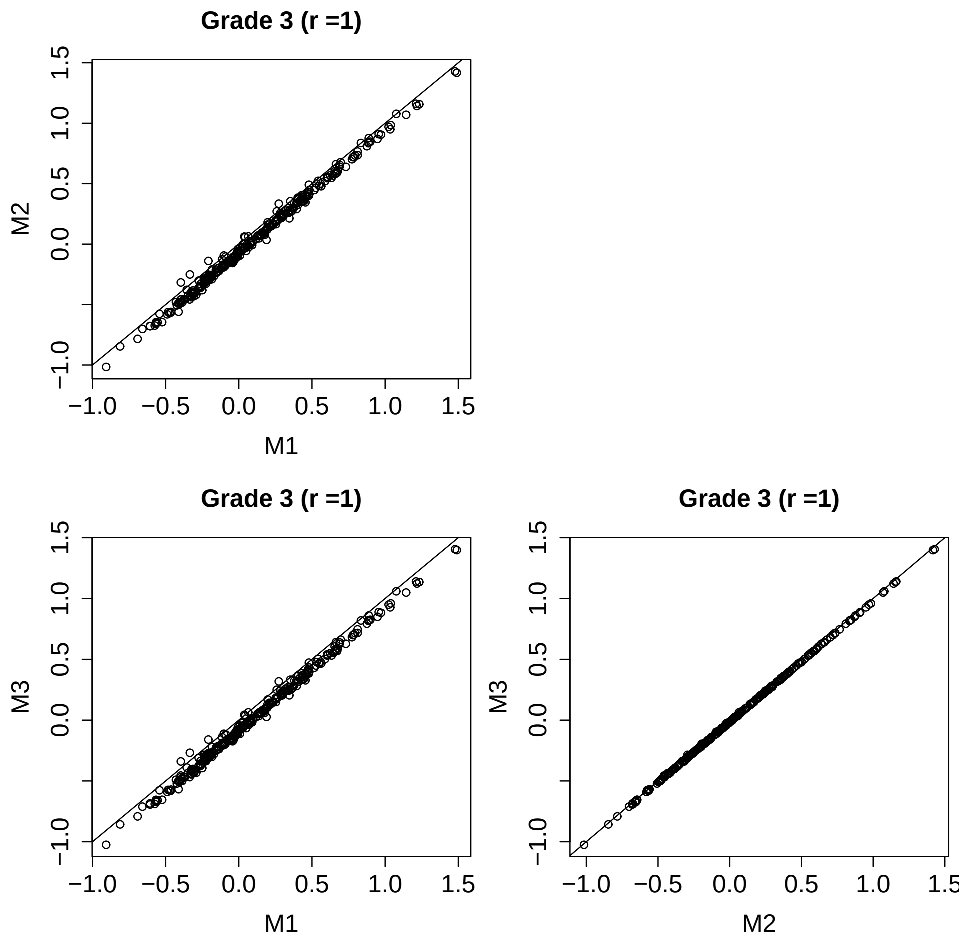
Missing Links Must Be Imputed

- **Bayesian method uses data augmentation to find posteriors given the observed data and priors**
- **Teacher links are required for augmentation**
- **Values of missing links matter for years prior to the last year of observed data**
  - **Persistence assumes prior teachers contribute directly to current year scores**
  - **After the last observed score, values of the missing links do not matter and there is no information abut these teachers in the data, we set these links to a teacher with zero effect**

Methods to Missing Links

- Method 1. Set all missing links to a teacher with zero effect
- Method 2. Give each student a "pseudo-teacher"
  - Pseudo-teacher links only to this one student
  - Estimate pseudo-teacher effects
  - Variance of pseudo-teachers effects is assumed to equal that of other teachers from the same grade
- Method 3. Give each student a "pseudo-teacher" but let pseudo-teachers have different variance components from other teachers

Teacher Effect Estimates Nearly Invariant to Method for Missing Link



RAND

Simulation Study of Sensitivity to Assumption That Scores are Missing At Random (MAR)

- **Generate data from persistence model with no covariates and no school effects**
- **All the α parameters equal one (i.e., the layered model)**
- $\epsilon_{it} = \sigma_t(\delta_i + \zeta_{it})$
  - $\delta_i \sim N(0, .7)$, $\nu^2 = .7$
  - $\zeta_{it} \sim N(0, .3)$
  - σ_t^2 **based on the marginal variance of our urban test data**

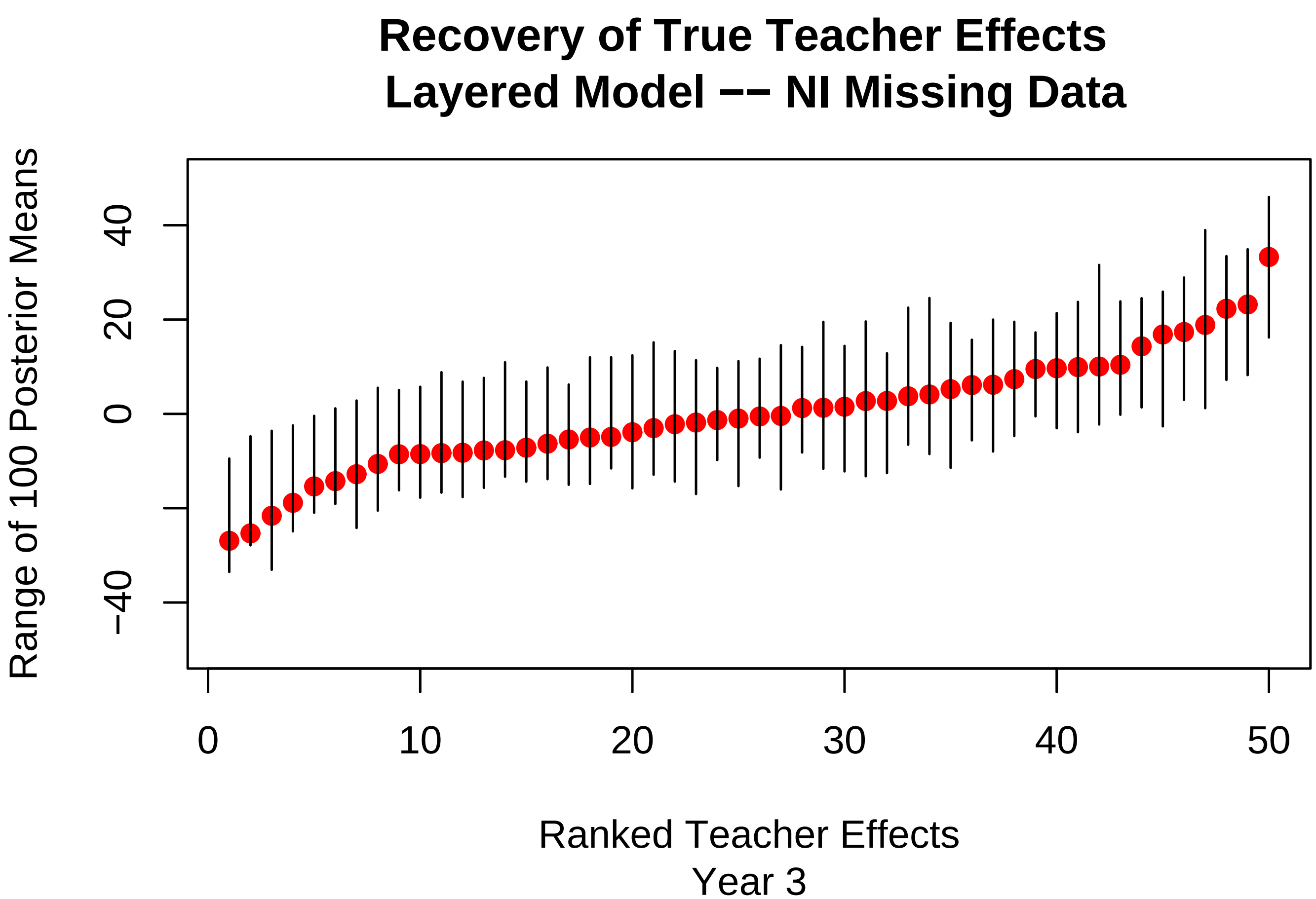
Simulation Study Missing Data Mechanism

- **Probability of missing depends on student's δ value**
  - **Number of observed year T_{obs} depends on δ through the model**

$$P(T_{obs} \geq t) = \frac{1}{1 + e^{\mu_t - \beta\delta}}, \text{for } t = 1, \ldots, 5 \ / \quad (1)$$

  - **$\beta > 0$ – students with lower values of δ are more likely to have more missing scores**
  - **Pattern of missing is selected at random**
  - **Probability of missing scores matches our urban test score data**
- **5 grades, 50 teachers per grade, 25 students per teacher**

Teacher Effects Are Relatively Robust to Violation of MAR Assumption

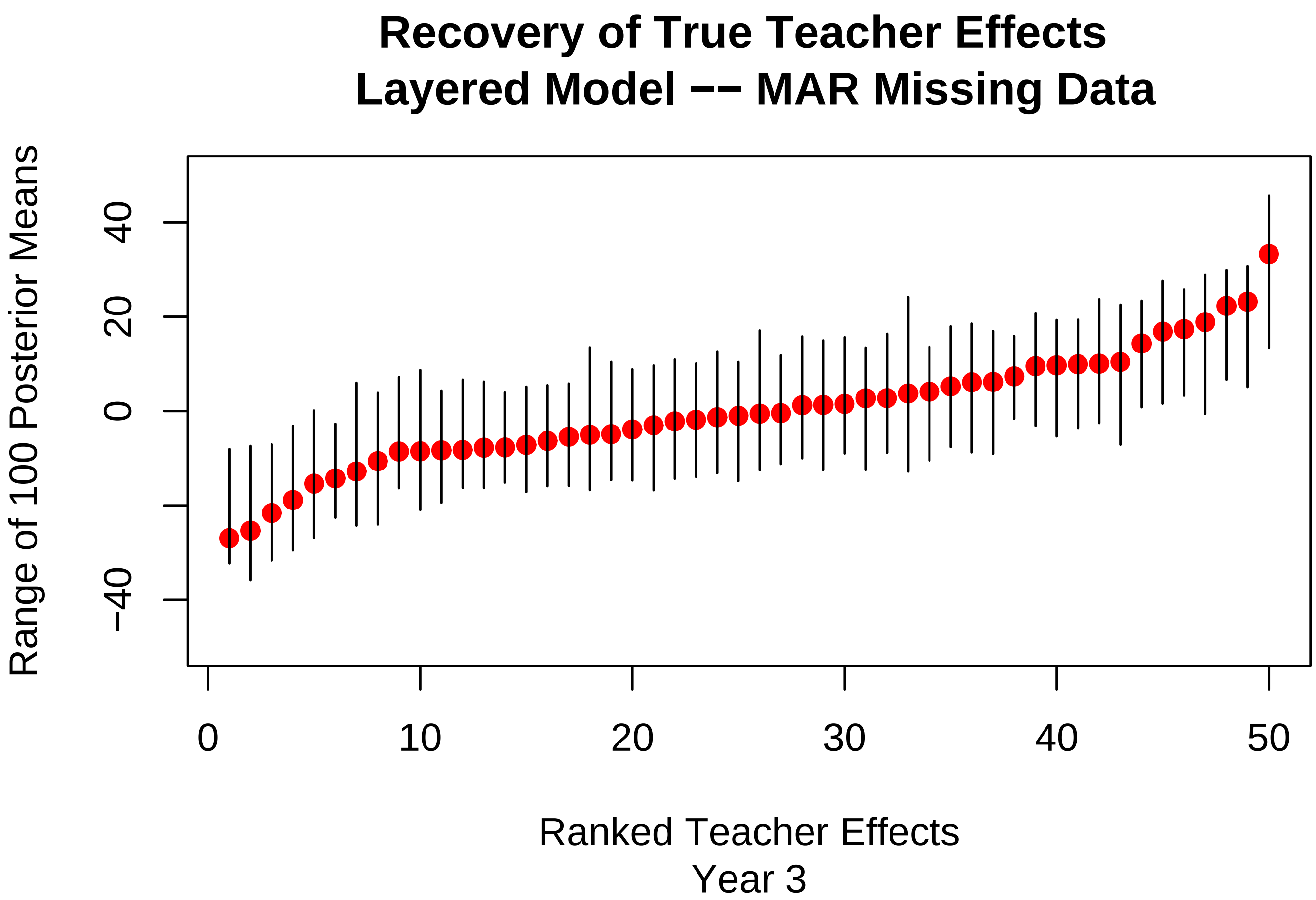


RAND

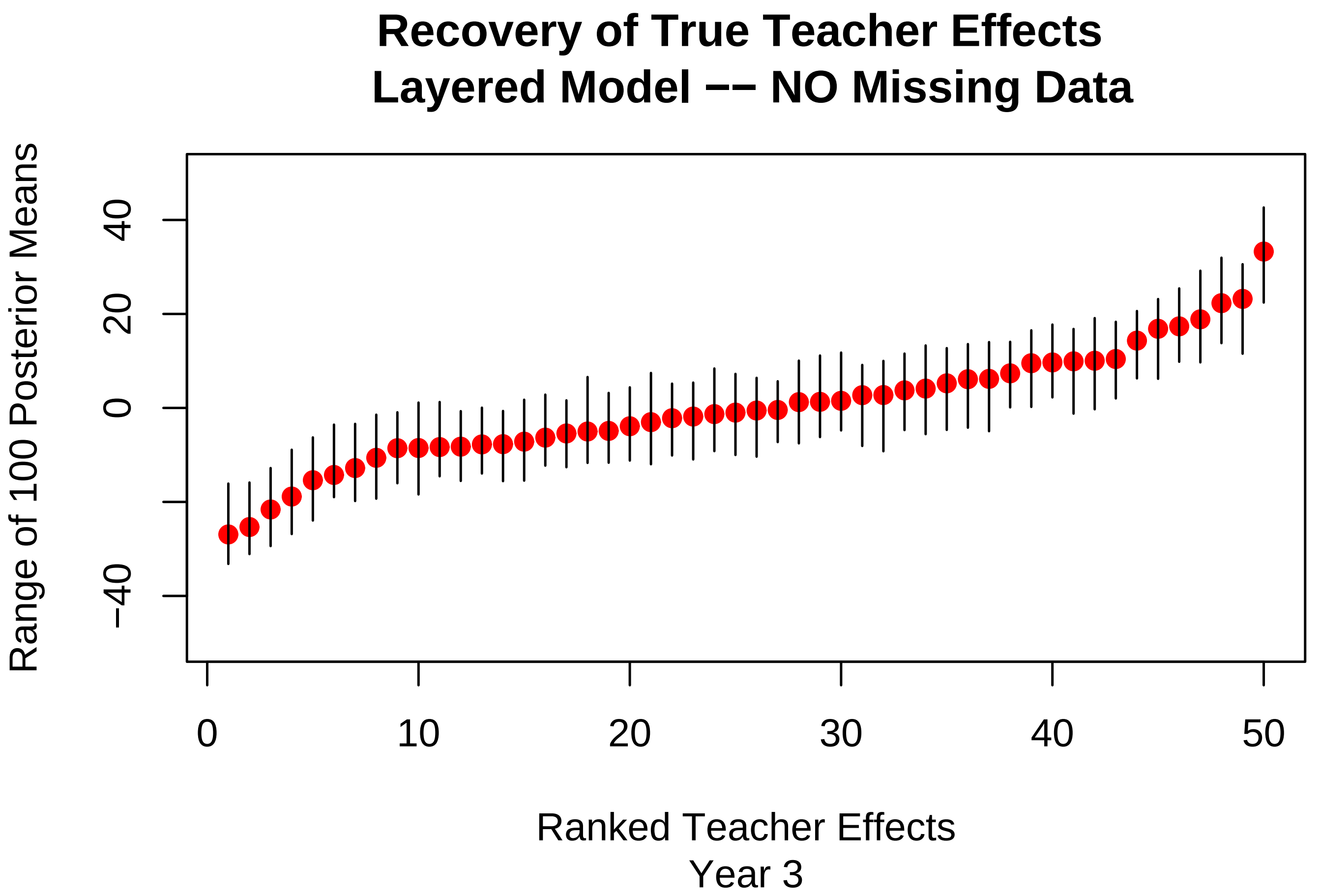
But There is Bias for Teachers at the Low End of the Sample and In the Overall Average Effect

- **Estimates tend to be above the true value for teachers at the low end of the sample**
- **Average estimated teacher effect estimate is greater than zero for grades 1 to 3**
  - **For grades 1 to 5, means are 34.1, 22.5, 9.5, .2, and -.1 % of standard deviation unit**
- **Bias can result in misclassifying teachers as above or below zero**
- **Bias occurs only when data are not MAR**
- **Bias occurs only when we are missing teacher links**

Teacher Effects When Data are MAR



Teacher Effects When Data are Complete



RAND

Outline

- Persistence of teacher effects
- Effects of student background variables
- Missing scores and missing teacher links
- **Modeling approach**

Are the Complex Models Worth the Effort?

- Persistence model, layered model, and cross-classified models lack transparency
- We investigate the estimators and compare average mean square error for several approaches.

Tractable Simple Case for Analytic Investigations

$$y_{i1} = \mu_1 + \boldsymbol{\phi}'_{i1}\boldsymbol{\theta}_1 + \epsilon_{i1}$$

$$y_{i2} = \mu_2 + \alpha_{21}\boldsymbol{\phi}'_{i1}\boldsymbol{\theta}_1 + \boldsymbol{\phi}'_{i2}\boldsymbol{\theta}_2 + \epsilon_{i2}$$

- **Complete data**
- **Known parameters**
- **No year one clustering**

$$y_{i1} = \mu_1 + \epsilon_{i1}$$

$$y_{i2} = \mu_2 + \theta_2 + \epsilon_{i2}$$

- **Σ equals the variance covariance matrix for (ϵ_{i1}, ϵ_{i2})**
- $d_i = y_{i2} - y_{i1}$

Seven Approaches to Modeling This Data

1. **Fixed effects using y_2**
2. **Fixed effects using d**
3. **Random effects modeling y_2 ignoring y_1**
4. **Random effects modeling d ignoring y_1**
5. **Random effects modeling y_2 and y_1 jointly**
6. **Random effects modeling d and y_1 jointly**
7. **Random effects with y_1 as a covariate in the model for y_2**

Teacher Effects Estimated by the BLUP

- **The BLUP, $\widehat{\theta}$, is given by**

$$\widehat{\boldsymbol{\theta}} = \left(\mathbf{Z}'\mathrm{R}^{-1}\mathbf{Z} + \mathrm{D}^{-1}\right)^{-1}\mathbf{Z}'\mathrm{R}^{-1}\mathbf{r}$$

  - R **– the block-diagonal covariance matrix for the entire vector of residual error terms**
  - $\mathbf{Z}$ **– the matrix of links from student scores to teachers**
  - D **– covariance matrix for the random teacher effects ($\tau\mathrm{I}$)**
  - r **– the vector of estimated residuals $\mathbf{y} - \boldsymbol{\mu}$**

Evaluating the BLUP

- Evaluating $\mathbf{Z}'\mathbf{R}^{-1}\mathbf{r}$
  - Elements are proportional to classroom means of adjusted residuals, e
  - Residuals determined by the $\mathbf{R}^{-1}$ for each approach
- Evaluating $\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1}\right)^{-1}$
  - The matrix is diagonal
  - Elements are proportional to a “shrinkage factor” λ
    - For random effects, $\lambda = \tau/(\tau + \nu^2)$
    - ν^2 is the variance of the adjusted residual mean conditional on θ
    - For fixed effects, $\lambda = 1$
- $\widehat{\theta} = \lambda\bar{e}$
- MSE = $\lambda\nu^2$

Joint Modeling Provides Smallest MSE of Seven Alternatives

Approach	MSE ($\lambda \times \nu^2$)
Fixed Effects with y_2	$1 \times \sigma_2^2/n$
Fixed Effects with d	$1 \times \sigma_d^2/n$
Random Effects, y_2 Alone	$\left(\frac{\tau}{\tau+\sigma_2^2/n}\right) \times \sigma_2^2/n$
Random Effects, d Alone	$\left(\frac{\tau}{\tau+\sigma_d^2/n}\right) \times \sigma_d^2/n$
Joint Modeling (y_1, y_2) or (y_1, d) y_1 as covariate	$\left(\frac{\tau}{\tau+\sigma_2^2(1-\rho^2)/n}\right) \times \sigma_2^2(1-\rho^2)/n$

With Clustering in Both Years Joint Modeling Adjusts for Year 1 Teachers

- **Step 1. Generate preliminary year 1 teacher effect estimates ignoring year 2 teacher assignments**
  - **Similar to Approach 3, estimate residuals from regressing estimates of ϵ_1 on ϵ_2, create classroom means of the residuals, and shrink toward zero**
- **Step 2. Adjust $r_1 = y_1 - \mu_1$ by the estimated year 1 teacher effects to obtain $\tilde{r}_1$**
- **Step 3. Adjust $r_2 = y_2 - \mu_2$ by $\beta_{21}\tilde{r}_1$**
- **Step 4. Average the adjusted year two residuals by classrooms and shrink**

Joint Modeling Is Preferable When Clustering Exits in Year 1

- **When class assignments are completely balanced, teacher effect estimators equal those that ignore clustering in year 1**
- **When not balanced, the estimator is not equivalent to using y_1 as covariate and should be more efficient because it adjusts for year 1 teachers and it uses correct adjustment for year 1 residuals**

What Have We Learned About the Persistence of Teacher Effects?

- Teacher effects dampen over time in a large school district
- Inferences can be sensitive to model assumptions about dampening
  - Persistence model finds more teachers distinct from zero
- Bayesian implementation improves computation and makes the model feasible even for large problems

What Have We Learned About the Effects of Student Background Variables?

- Linear covariate adjustments can create bias by over adjusting
- We cannot assume joint modeling of multiple scores will remove effects of student background variables
  - Particularly concerned when combining segregated populations where gains might differ across groups
- Best approach might be to compare teachers teaching similar populations

What Have We Learned About Missing Data?

- Missing teacher links are a challenge to modeling incomplete data
- Ordering of teachers appears robust to approach to missing links and to violations of MAR
- Inferences about teachers being above or below average might be sensitive to violations of MAR
  - Centering estimated effects might provide an ad hoc fix to bias
- Further studies should:
  - Explore alternative teacher assignment mechanisms
  - Explore missing data mechanisms that depend on gains
  - Explore the effect of missing data when αs are less than one

What Have We Learned About Complex Models?

- Joint modeling is doing regression adjustments
  - In simple special cases joint modeling is identical to covariate adjustment with year one score
- Joint modeling of scores is likely to be more efficient than alternatives
  - Uses prior year score efficiently and adjusts for other year teachers
- More work needs to be done to explore gains in efficiency from joint modeling
- More work needs to be done to see if gains for joint modeling hold when the data are very incomplete and assumptions are violated