RAND EDUCATION

From Data to Bonuses – A Case Study of Awarding Teachers Pay on the Basis Of Their Students' Progress

Daniel F. McCaffrey, Bing Han, J.R. Lockwood

February 29, 2008

A Pay for Performance System

- Use administrative data to collect information on student achievement and student-teacher links
- Use value-added methods to estimate teacher performance
- Award bonuses on the basis of the value-added measures
- Each step requires many decisions and currently little guidance exists for making them

Outline for Talk

- Preparing data for estimating teacher effects
  - Selecting teachers
  - Selecting students
  - Data quality
- Case study data
- Comparison of value-added performance measures
- Comparison of decision rules for awarding bonuses

Selecting Teachers for Evaluation

- Tested grade-levels and subjects determine the teaching tasks to be evaluated (e.g., grade 8 mathematics or grade 5 reading)
- For selected teachers, evaluated teaching task must constitute a sufficiently large portion of a teacher's responsibility to be used in pay determination
- Selected teachers must be appropriate to include in the peer group used to determine other teachers relative performance
  - Most value-added measures of performance are relative to the average performance of a group of teachers
  - Decision rules for bonuses often depend on ensemble of teachers included in the evaluation sample

Issues with Teacher Selection

- Special education classes
- Small classes
- Multiple grade-levels
- Identifying courses
  - The relevance of some courses to tested subjects is ambiguous
  - Drama or speech courses are taught in middle schools, but are they relevant to English Language Arts tests?
  - Is reading relevant to English Language Arts tests? Is English relevant to reading? Is social studies relevant to writing?

Selecting Students

- For which students should a teacher be held accountable?
  - How much time is required in a teacher's class for a student's performance to be attributable to the teacher?
  - How do we apportion attribution to teachers when a student received instruction on a subject from multiple teachers?
    - Simultaneous instruction in multiple related courses or from regular and special education teachers
    - Sequential instruction
    - Administrative data do not always provide full details or timing of courses
    - Is the proportion of instructional time the correct way to apportion courses to teachers?
    - Can such apportioning of time be implemented in performance measurement?

Other Data Issues

- Administrative data can be incomplete or not up to date
  - Many teachers found errors in rosters created from administrative data maintained by the district
- Data from different administrative databases can be contradictory
- Data can contain multiple records for students from a single year
- Common data cleaning methods used for research might not suffice when preparing data for determining compensation
  - Lost cases might signal to teachers that some students do not count and lead to negative consequences
  - Teachers might request full accounting; unverifiable data fixes could result in challenges to compensation decisions and weaken confidence in the system

Case Study Samples

- Large urban school district
- 50% African-American, 36% White, 11% Hispanic, and 3%Asian or other ethnic group
- Sample includes 37,887 students enrolled the district's middle schools during at least part of one or more of the 2004-05, 2005-06, or 2006-07 school years
- Teacher sample is all teachers who taught mathematics to these cohorts of students in middle school during the 2005-06 and 2006-07 school years
  - Includes classes regardless of size
  - Includes special and regular education classes
- n=478 in 2006-07 and 476 in 2005-06, 338 teachers in both years

Case Study Data

- Three source files:
  - Enrollment files
  - Course files
  - Test files
- Student scores on the state mathematics, reading/English langauge arts, science, social studies tests for grades 3 to 8
  - Use both scale scores and rank-based z-scores
- Student grade-level and background variables including race, gender, special education status
- Teacher scores on items from Learning Mathematics for Teaching Project's Multiple Choice Measures of Mathematical Knowledge for Teaching (LMT)

Performance Measures

- 24 performance measures
  - Factorial crossing of method, statistical adjustment (with or without shrinkage), and test scale (raw scale scores or z-scores)
- The 8 methods used are:
  - ANCOVA, Regression Residuals, Lookup-Tables
  - Average gain scores
  - Multivariate ANCOVA
  - Multivariate mixed effects models
    - Variable persistence and layered (complete persistence) models
  - Fixed Effects
- Shrinkage multiplies an estimate by a factor less than one chosen to minimize the average squared error between estimates and true values by reducing noise

Focus on Four Methods

- **Two simple, transparent methods with intuitive appeal to stakeholders**
  - **1. ANCOVA**
  - **2. Gain Scores**
- **Two complex methods more widely accepted by methodologists**
  - **3. Multivariate mixed models**
  - **4. Fixed effects**

1. The ANCOVA Method

- Based on traditional analysis of covariance approach to estimating adjusted group means
  - Use linear regression on a single prior year mathematics score to control for student inputs
- Must parameterize model so effects to sum to zero to assure stability across years
- Comparison uses raw scores, without shrinkage

2. The Gain Score Method

- Gain scores equal current year score less the prior year score
- Performance measure equals the simple average of the teacher's students' gain scores
- Requires scores on a common scale across grades
- Comparison uses raw scores, without shrinkage

3. The Multivariate Mixed Model Method

- Model the joint distribution of each student's vector of scores
- Explicitly models each year's score as a function of the contribution of the current and past year teachers and student-specific residual errors
- Teacher effects are modeled as random variables from common distribution
- Student residuals have unspecified correlation structure and variance depends on grade-level
- Explicitly models the persistence of teacher's effect on students' future outcomes
- Fit separate models for each cohort
- Fit variable persistence model using Bayesian framework
- Fit to z-scores and method implicitly provides shrunken estimates

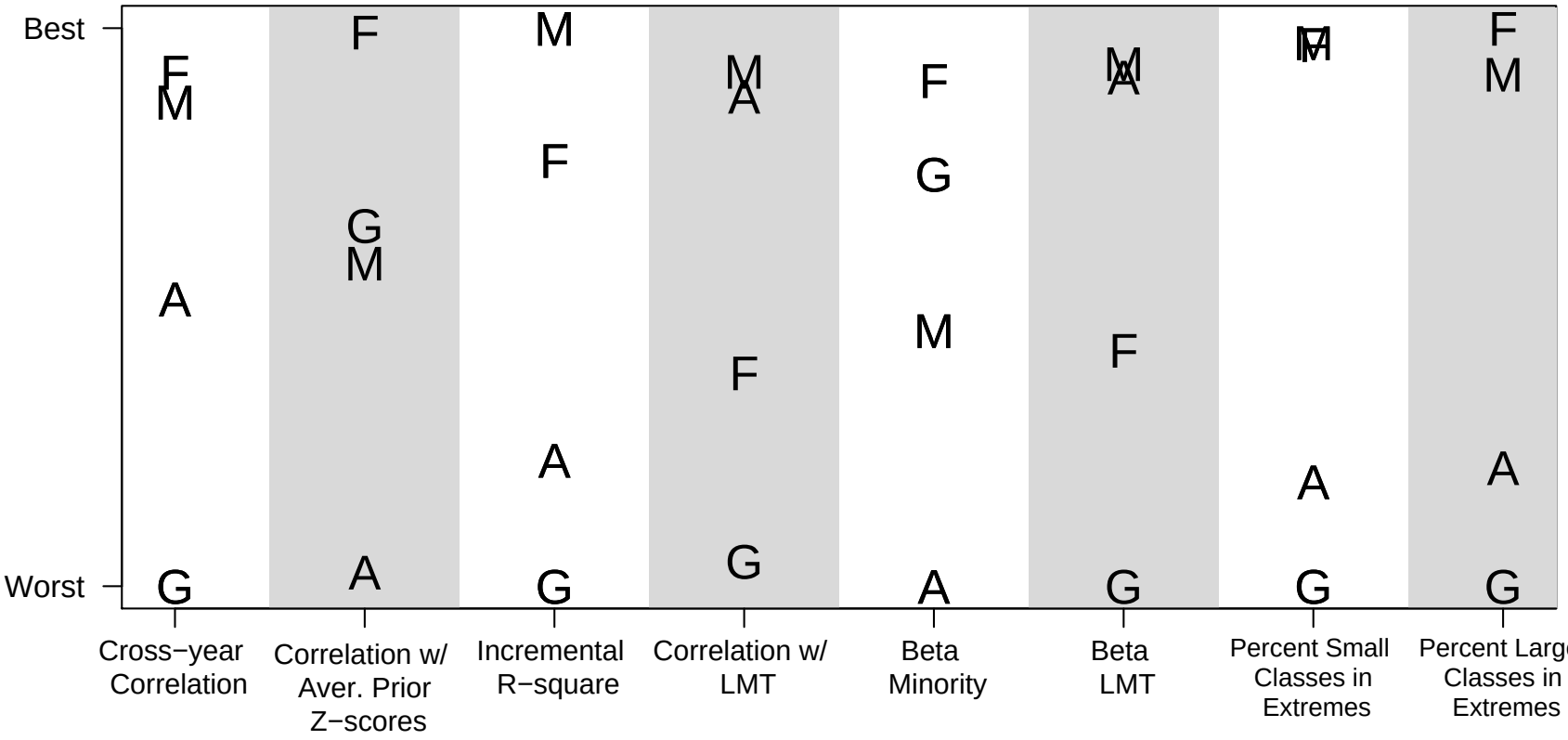
4. The Fixed Effects Method

- Model for the joint distribution of each student's vector of scores
- Include indicator variables for every student
- Include indicator variables for every teacher
- Must parameterize teacher effects to sum to zero
- Uses within-student variation to estimate teacher effects
- Scores must be on a scale where a constant additive student effect is plausible
- Assumes teacher effects last only one year
- Fit separate models for each cohort
- Comparison uses z-scores and shrinkage

Sources of Variance in Estimated Performance Measures

- **Signal**
- **Bias**
- **Noise**

Properties of Selected Performance Measures



Simple Methods Have Poor Properties, Complex Methods Have Better Properties

- ANCOVA method has large bias favoring teachers of students with higher prior achievement
- Gain score method has large noise
  - Very unstable across years
- Mixed models have small bias and relatively little noise
  - Strongest correlation with LMT
  - Cross-year correlation greater than 0.50
  - Weakly favors teachers of students with higher prior achievement
- Fixed effects have even less bias and relatively little noise
  - Cross-year correlation of greater than 0.50
  - Very low correlation with prior achievement and minority status
  - Might weakly favor teachers of students with lower prior achievement

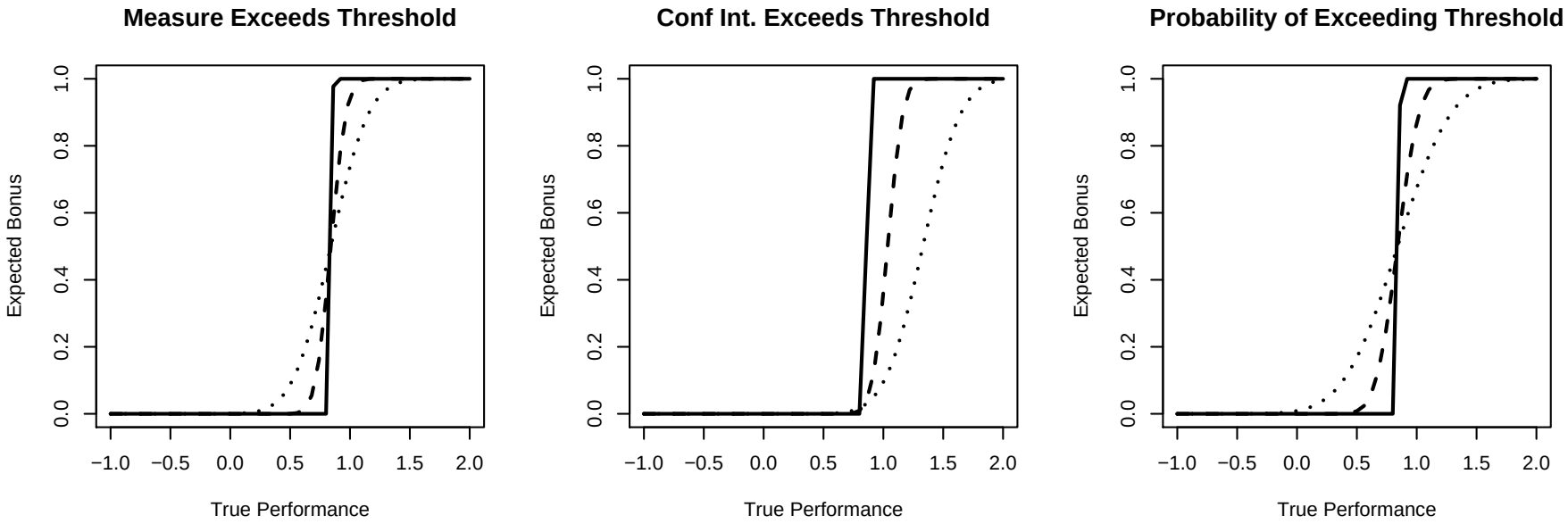
Comparison of Significantly Better than Average Performers

- Compared with the other method ANCOVA finds many more teachers' performance classified as significantly better than average
  - Teachers of classes with high prior achievement and low percent minority
  - Discrepancies with other methods persist across years
- Fixed effects and multivariate mixed models have high agreement
  - Performance measures correlate .86
  - Agree for 90% of teachers classified as significantly above average performance or not
- But when they disagree
  - Mixed models favor teachers with above average LMT scores, teaching high achieving students in classes with low percent minority students
  - Fixed effects favor teachers with below average LMT scores, teaching lower achieving students in classes with high percent minority students
  - Discrepancies do not persist across years

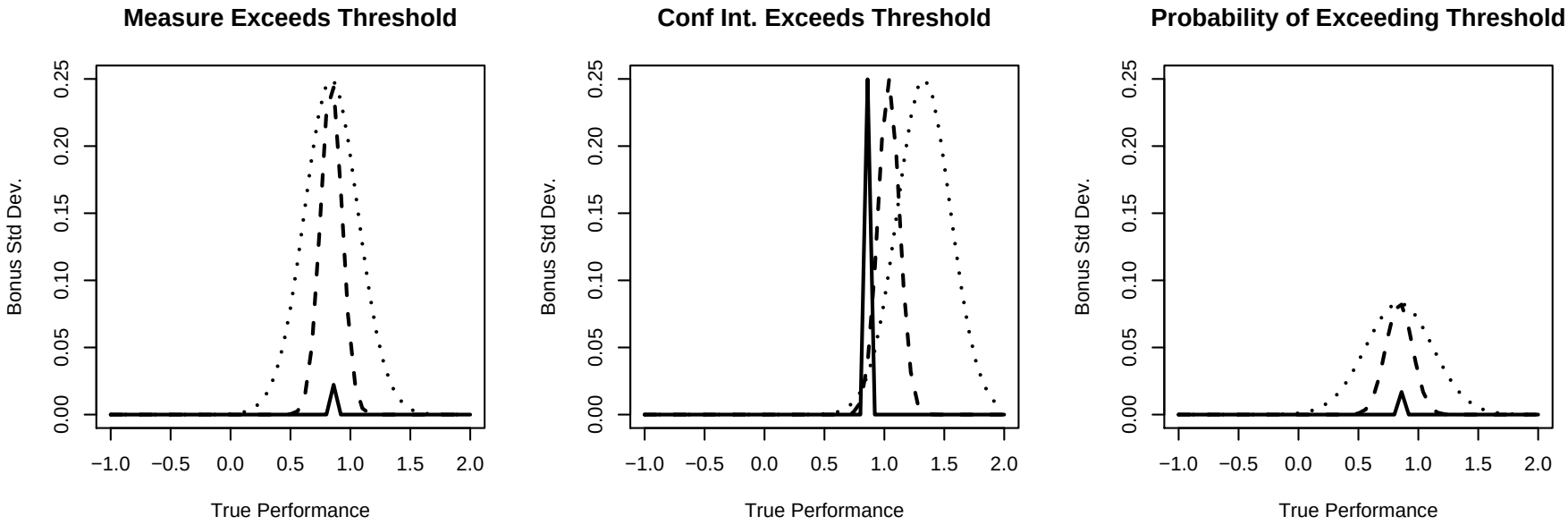
Study of Bonus Decision Rules

- Conducted analytic investigation of bonuses as a function of
  - Decision rule
  - Noise
  - Bias
- Conducted an empirical study using alternative decision rules with each performance measure

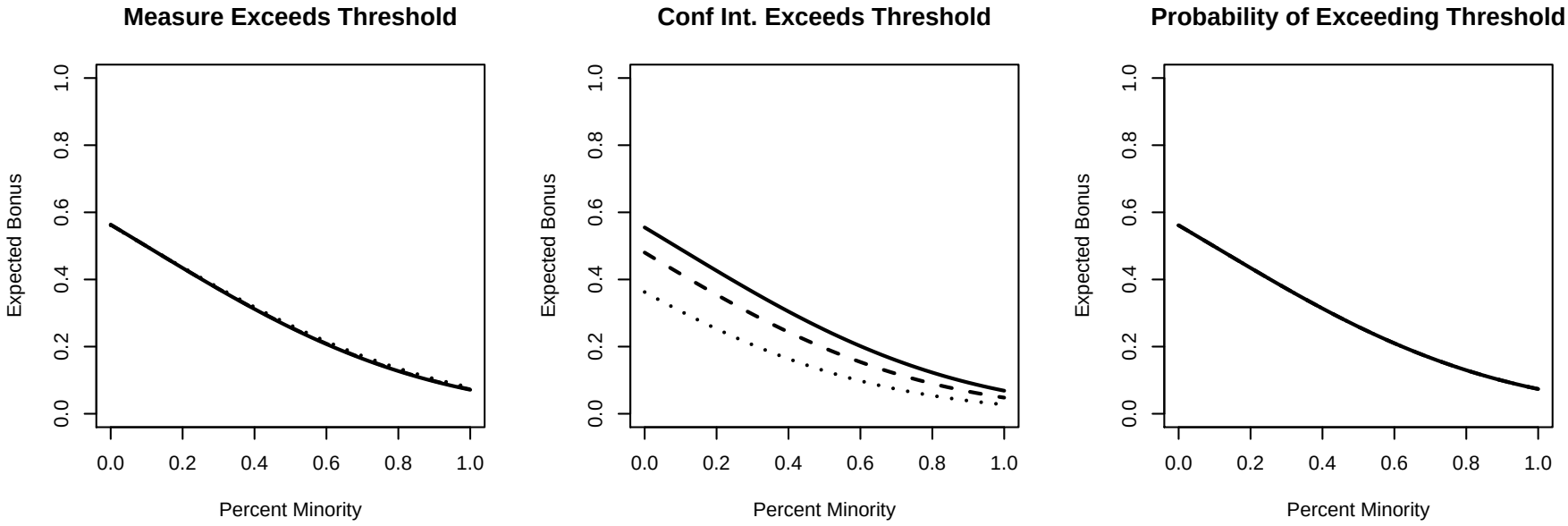
Expected Bonuses Vary with Decision Rule



Uncertainty in Bonuses Varies with Decision Rule



Bias in Performance Measure Can Greatly Change Expected Bonuses for Teachers with Truly Equal Performance



Awarding Pay on the Basis of Student Performance Is a Complex Process

- **Requires extensive data preparation**
  - **Selection of teacher**
  - **Selection of students**
  - **Cleaning and verifying administrative data**

Awarding Pay on the Basis of Student Performance Is a Complex Process

- **Performance measures are not all equal**
  - **ANCOVA method has large bias**
  - **Gain score method has large noise**
  - **Mixed models and fixed effects have good properties**

Awarding Pay on the Basis of Student Performance Is a Complex Process

- **Decision rules susceptible to noise and bias**
  - **Greater noise results in greater awards to low performing teachers and smaller awards to high performing teachers**
  - **Bias makes expected bonuses different for equal-performing teachers depending on the students in their classes**
  - **Different decision rules correspond to different evaluation of the costs of Type I and Type II errors**

Choosing Systems Will Require More Information on Teachers

- How will teachers with various true levels of performance and teaching in different contexts respond to different levels of expected bonuses and different levels of uncertainty in bonuses?
- How will these responses interact with other features of performance measures and bonus decision rules?
  - Will teachers' responses to the payout distribution depend on how statistically complex or transparent the performance measures are?
  - Will teachers' responses depend on how the distribution of bonuses varies across teachers of different types of students?
  - Will teachers responses depend on how the information about performance is presented to them?
- We need to identify leverage points in the system and develop experiments to find out teachers' responses to these features
- We can then design the system to obtain the responses we desire