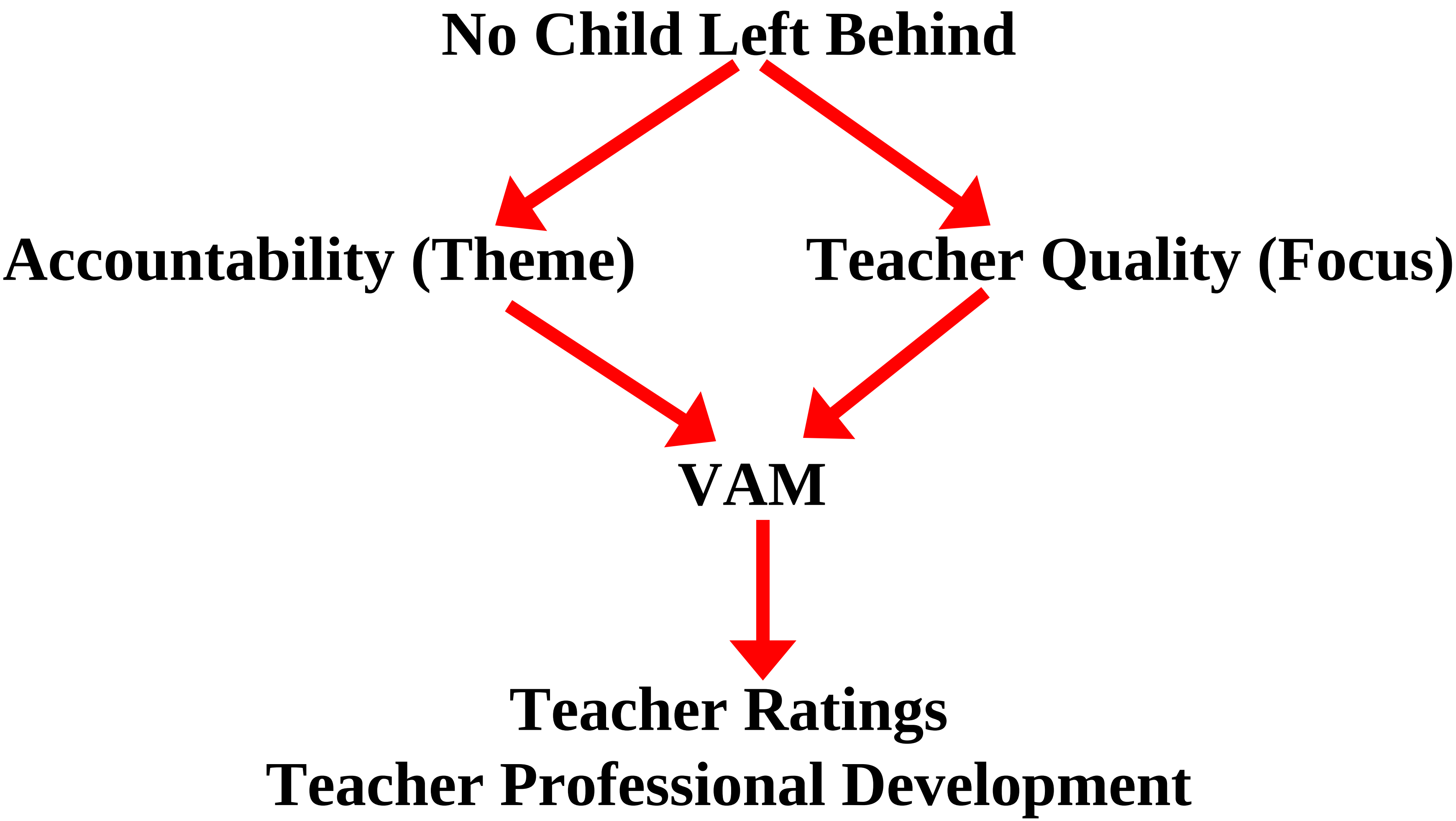
Value-added Modeling:
What does “Due Diligence” Require?

Henry I. Braun
Educational Testing Service

Presented at the University of Maryland
October 21, 2004

Introduction



No Child Left Behind

Accountability (Theme)

Teacher Quality (Focus)

VAM

Teacher Ratings

Teacher Professional Development

The Logic of VAM

If good teaching is critical to student learning, then can’t student learning (or its absence) tell us something about the quality of teaching?

A VAM Claim

Using sophisticated statistical methods, it is possible to objectively isolate the contribution of each teacher to student learning.

Early Concerns

TVAAS is a “black box” and it’s too complicated to use and to explain.

The response...and a rejoinder.

Goals of this Presentation

A. Discussion of “Due Diligence”.

B. Outline of some forthcoming research studies.

C. Discussion of the convergence of policy and methodology.

A. Due Diligence

$$y_t^k = b_t^k + u_t^k + e_t^k \qquad (1)$$

$$y_{t+1}^{k+1} = b_{t+1}^{k+1} + u_t^k + u_{t+1}^{k+1} + e_{t+1}^{k+1} \qquad (2)$$

where

y_t^k = student score in grade *k*, year *t*

b_t^k = district mean score in grade *k*, year *t*

u_t^k = contribution of the teacher in grade *k*, year *t*

e_t^k = unexplained variation in student score in grade *k*, year *t*

- Equations (1) and (2) are treated as a mixed model, with the *b* coefficients estimated as fixed effects and the *u* coefficients estimated as random effects.
- Models for subsequent years follow the same pattern. In the aggregate, such a set of equations is referred to as a “layered model.”
- Can also include data from other subjects.
- Using a layered model brings more data to the estimation of teacher effects—but at the cost of making more assumptions about the processes generating the data.

Assumption	A priori Reasonableness	Empirical Sensitivity
1. Construct validity of test scores	Varies by state and by subject within states	No information
2. Interval scale property	Limited. Probabably okay locally.	Depends on vertical scaling procedure.
3. Negligible selection bias for gains analysis	Low	Moderate to High (Intro of student covariates should help somewhat).
4. Missing data MAR	Low	No information.
5. Linear mixed model	Moderate	Moderate (Compare with fixed effects formulat-ion).
6. Persistence of teacher effects	Low	Moderate to high.
7. No teacher x student score interaction	Moderate. Varies by teacher.	No information.

Outstanding Question

“Under what circumstances is it reasonable to interpret estimated teacher effects as measures of teacher effectiveness?”

i.e. When are we justified in making causal inferences from an observational study subject to (strong) selection bias?

Research Agenda

- Clarify meaning of “teacher effectiveness”.
- Investigate sensitivity of estimated teacher effects to departures from assumptions.
- Make progress in disentangling “true” teacher effectiveness from school and context contributions to student learning.
- Characterize settings where use of VAM strongly discouraged.
- Obtain external validation of estimated teacher effects as approximate measures of teacher effectiveness.
- Explore cost/benefit of using VAM for teacher evaluation.

B. Research Studies

i. Score scale properties

Assumption of a single interval scale across years justifies the aggregation of gains across years.

Need to look at typical gains as a function of grade and starting point.

Map trajections of students over time. (Perhaps disaggregated by race/gender).

Alternate formulation in terms of Markov transition matrices for single year gains.

Final Score Category

Initial Score Category	1	2	3	4
1				
2				
3				
4				

Note: Doesn’t require two tests to be on the same scale.

Let $P_{ij} = \Pr\left\{i \to j\right\}$ [$\hat{P}_{ij}$ based on data from total population in the grade]

$\left\{n_{i\cdot}^{k}\right\}$ and $\left\{n_{\cdot j}^{k}\right\}$ be observed marginal counts in class of teacher *k*.

Then

$e_{ij}^{k} = \hat{P}_{ij} \times n_{i\cdot}^{k} =$ expected number of students in cell (*i*, *j*) for teacher *k*.

$\left\{d_{ij}^{k} = n_{ij}^{k} - e_{ij}^{k}\right\} =$ differences between observed and expected for teacher *k*.

Can cumulate $\left\{d_{ij}\right\}$ over cohorts.

Apply to data from:

Tennessee STAR

- Randomized experiment
- Norm-referenced tests

ECLS-K

- Ordinary school settings
- Specially constructed developmental scale

State ??

- Ordinary school settings
- Norm-referenced or criterion-referenced tests

Can focus on particular subsets of $\left\{d_{ij}\right\}$

Can compare teachers with similar distributions of $\left\{n_{i\cdot} / n_{\cdot\cdot}\right\}$

Can group teachers by distributions of students' baseline test scores and compare estimated teacher effects by group.

With sufficient data, can compare results with those of "standard" VAM.

Methodological Issues

- Stable estimation
- Accommodating student covariates
- Incorporating student data from previous or following years.

ii. Missing Data

There can be substantial missing data in a database: students lacking test scores and/or teacher links in one or more years.

Carry out simulations to assess sensitivity of the VAM-based estimates to assumptions of missing at random (MAR)

Use existing data as a basis for creating a complete data set.

Develop one or more models for departures from MAR.

Complete data can be used to generate incomplete data sets with different degrees of departure from MAR.

Simulation Design

$$\begin{tikzcd}
D'' \arrow[d, "n\text{MAR}"'] & R \text{ [original data set]} \arrow[l, "\text{MAR}"'] \arrow[r, "\text{MAR}"] \arrow[rd, "n\text{MAR}"'] & D'_{1.0} \\
\left\{ R_{ij} \right\} & & \begin{bmatrix} D'_{0.9} \\ D'_{0.8} \\ \vdots \end{bmatrix}
\end{tikzcd}$$

Key:

D: denotes a completed data set

R: denotes a data set with missing data

R_{ij} : *i* indexes fraction of missing data

j indexes degree of departure from MAR

iii. External “Validation”

Relating standard and modified VAM estimates of teacher effects to:

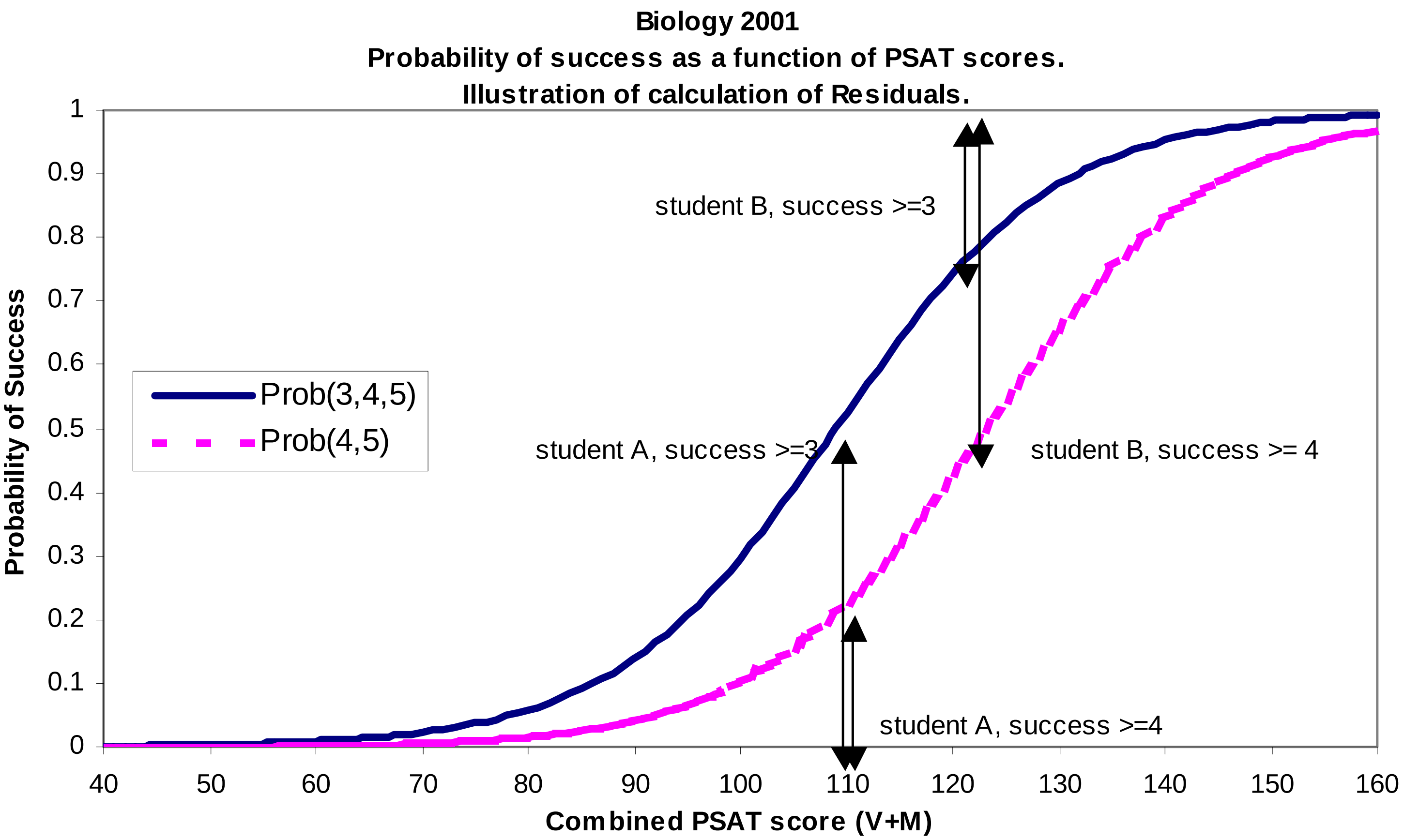
- Results of STAR experiment
- Results of teacher tests (NBPTS, PRAXIS I and PRAXIS II)
- Teachers’ pedagogy

A “Crude” Value-added Study

Conducted a national survey of teachers of AP Biology and AP U. S. History

Goals:

1. Describe current practice
2. Explore relationships among student outcomes and Teaching context and Teacher practice.



Student A had combined PSAT scores of 110 and received a 2 on the AP exam, a "failure". The residual is 0-.2259 = -.2259 for success >=4, and 0-.5245 = -.5245 for success >= 3. Student B had combined PSAT scores of 120 and received a 4 on the exam, a "success". Student B's residual for success 3 is 1-.7431 = .2569, andfor success >=4 is 1-.4278=.5722.

Used “Mean Class Residuals (MCR)” as a crude value-added measure.

Regressed MCR3 and MCR4 separately on:

- Cumulative school average PSAT score
- School and class context variables
- Teacher instructional practices

Preliminary Results

1. Cumulative school average PSAT score is the most powerful predictor.
2. Some school and class context variables significant.
3. A few pedagogical variables significant.

iv. Selection Bias

Are there ways of capitalizing on longitudinal structure of student data to get a handle on degree of selection bias, at least with respect to some aspect of the process by which students and teachers are matched?

C. Policy and Methodology

The challenge to the measurement community (and methodologists, in general) goes far beyond current concerns about the use of VAM for teacher evaluation:

How do we contribute constructively to policy-making, recognizing that the results of quantitative analysis are only one factor in the decision process?

Realities of Teacher Accountability

- Quantitative teacher evaluation is here to stay—and becoming more widespread.
- Inferences based on some value-added approach are very appealing and generally (much?) preferred to analyses based on student final status.
- Highly unlikely that we can ever unbiasedly estimate (average) teacher effectiveness.

Arguing that the models are not exactly right is not enough.

> It is not enough to admit that the model is subject – or even likely – to be found wanting in the future. We must be prepared to use many models, and find their use helpful for many specific purposes, when we already know that they are wrong – and in what ways. The model of a gas as a collection of hard round spheres undergoing mechanical collisions is demonstrably wrong in many ways. Yet it still serves us well in thinking about certain phenomena. In data analysis ...we must be quite explicit about the deficiencies of the models with which we work. If we take them at face value, we can – all too frequently – be led to unreasonable and unhelpful actions. If we try to make them "fit the facts," we can ensure sufficient mathematical complexity to keep us from any useful guidance.
>
> (Tukey, 1963)

Three-pronged Approach

1. Continue empirical investigations, especially sensitivity analyses

2. Evaluate different uses of VAM results in real world settings
    - Triage for identifying teachers in need of support
    - Rubin, Stuart, & Zanutto suggestion.

3. Communicate issues to non-technical audiences

With education continuing to be a high-profile political issue, the measurement community should engage in a more intense discussion of how we should interact with policy-makers, education administrators and the public at large.