Essential Elements of a System to Measure the Performance of Schools and Classrooms/Teachers with Respect to Student Achievement

Three criteria:

1. **Criterion validity/alignment** Are the indicators measured in terms of student outcomes valued by students and society?
2. **Statistical:** Are the indictors accurate in the sense of measuring true school or classroom productivity, as opposed to other non-school factors that contribute to student achievement?
3. **Behavioral**: Are the indicators non-corruptible?

Punch Line:

Appropriately designed value-added models satisfy the statistical criterion and generally satisfy the non-corruptibility criterion.

Behavioral: Are the VA indicators non-corruptible?

1. Do VA indicators provide proper incentives to make decisions to maximize growth of student achievement for all students? For example, a proficiency rate indicator provides an incentive to focus resources only on students close to the proficiency standard (cut point).
2. Are all students included?
3. Are control variables fixed or subject to manipulation by schools? (Example: gender and race/ethnicity are largely fixed student attributes, whereas participation in special education is both a student atttribute and a factor that is partly determined by educational staff.)
4. Are new test forms used for each test administration, so that students and staff are not familiar with test items used on previous assessments?

Statistical Criteria

1. **Validity**
2. **Reliability**
3. **Minimum Total Error**

1. **Validity:** Is the indicator derived from a statistical model that captures the true contribution of a school, classroom/teacher, or other educational unit to growth in student achievement, and “filters” out the non-school factors that contribute to student achievement, including differences in the types of students served by different schools?
   a. We refer to a model that meets this criterion as a value-added model (VAM).
   b. We refer to an error due to imperfect model design as a model error. An example of a model error is selection/assignment bias: the bias in measuring school performance, if any, due to failure to control for differences across schools in the types of students served by those schools.
   c. Average achievement or a proficiency rate is an example of an indicator that, if used to measure school performance, generally exhibits large selection bias.

2. **Reliability:** Is the indicator precise (low statistical/sampling error) in the sense that the estimated performance indicator is unlikely to differ substantially from the performance indicator that is defined by the statistical model?
    a. If not, measured performance is likely to be highly unstable from year to year.
    b. In general, reliability depends on two factors:
        i. the number of students used to produce an indicator for a given school, grade level, or classroom.
        ii. The complexity of the model – greater complexity may reduce precision.

3. **Minimum Total Error:** Are all sources of error – model errors and sampling error – low?
   a. Technical criterion: minimize mean squared error from all sources (MSE), if there are multiple sources of error.

Example

NAEP Mathematics Examination Data

(A)

Average Test Scores by Year

Grade	1973	1978	1982	1986
3rd	219.1	218.6	219.0	221.7
7th	266.0	264.1	268.6	269.0
11th	304.4	300.4	298.5	302.0

NAEP Mathematics Examination Data

(B)

Average Test Score Gain From Year to Year for Each Cohort

Grade	73 to 78	78 to 82	82 to 86
3rd to 7th	45.0	50.0	50.0
7th to 11th	34.4	34.4	33.4

Source: Dossey et al. (1988).

Model I. A Value-Added Model of School Performance for a Given Subject, Grade, and Year – A T2 Model

$$Y_{1i} = \zeta + \lambda Y_{0i} + \beta' X_i + \sum_{k(\text{school})} \alpha_{1k} S_{1ik} + \varepsilon_{1i}$$

Posttest

Pretest

Student Characteristics

α_{1k} = School Effect

School Participation Indicator

Stud. Error Term

λ = Post-on-Pre Link

Model II. A Value-Added Model of School and Classroom/Teacher Performance for a Given Subject, Grade, and Year – A T2 Model

$$Y_{1i} = \zeta + \lambda Y_{0i} + \pi' X_i +$$

$$\sum_{k(\text{school})} \alpha_{1k} S_{1ik} + \sum_{k(\text{school})} \sum_{j(\text{classroom})} \beta_{1jk} C_{1ik} + \varepsilon_{1i}$$

Model III. A "T3" Value-Added Model of School Performance

$$Y_{1i} = \zeta_1 + \lambda Y_{0i} + \beta' X_i + \sum_{k(\text{school})} \alpha_{1k} S_{1ik} + \varepsilon_{1i}$$

$$Y_{2i} = \zeta_2 + \lambda Y_{1i} + \beta' X_i + \sum_{k(\text{school})} \alpha_{2k} S_{2ik} + \varepsilon_{2i}$$

Are there components of student achievement growth that are not directly included in the model? Capture these components indirectly by linking growth equations (for the same individuals) across years, as indicated below:

$$\varepsilon_{1i} = \gamma_i + e_{1i}$$

$$\varepsilon_{2i} = \gamma_i + e_{2i}$$

Recommendations for Building a Value-Added System

Categories:

1. Data Requirements and Data Quality
2. Model Design
3. Using Value-Added Analysis to Evaluate Programs and Policies
4. Alignment with School, District, and State Policies and Practices

VA Recommendation #1.

Use all available longitudinal data to estimate a "T3+" model; that is, use a model that exploits repeated observations on students to control for student selectivity.

Examples of T3+ Models:

- TVAAS Layered Model (Sanders and colleague)
- Multivariate Value-Added Assessment (Lockwood and McCaffrey and colleagues)
- Generalized Value-Added Model with Conditional Random Effects And Multivariate Shrinkage (CRE-MS) (Meyer)

But…

How do you estimate school performance using data from the first two grades that are tested (typically 3rd and 4th grade)?

1. You could wait a year (until a student is in 5th grade) to produce performance estimates for 4th grade.
2. You could estimate the best possible "T2" model.

VA Recommendation #2.

Include explicit measures of student characteristics in the model if: (a) the number of longitudinal observations per student is limited (so that indirect control for differences in student growth trajectories is limited) or (b) you want to maximize control for student differences across schools.

VA Recommendation #3.

If using prior test scores as regressors, control for test measurement error.

VA Recommendation #4.

Extend the model to allow for local conditions, as required. Example: mid-year testing.

Model Design References

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). *Journal of Educational and Behavioral Statistics, 32*(2), 125–150.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101.

Meyer, R. H. (1992). *Applied versus traditional mathematics: New econometric models of the contribution of high school courses to mathematics proficiency* (Discussion Paper No. 966-92). Madison: University of Wisconsin–Madison, Institute for Research on Poverty.

Meyer, R. H. (2008, April). *A generalized value-added model with conditional random effects and multivariate shrinkage*. Paper presented at the National Conference on Value-Added Modeling, Madison, WI.

Meyer, R. H., & Christian, M. S. (2008, February). *Value-added and other methods for measuring school performance: An analysis of performance measurement strategies in teacher incentive fund proposals.* Paper prepared for the conference on Performance Incentives: Their Growing Impact on American K-12 Education, Nashville, TN.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*, 299–311.

Sanders, W., Saxton, A., & Horn, B. (1997). The Tennessee Value-Added Assessment System: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluational measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press, Inc.

Webster, W. J., & Mendro, R. L. (1997). The Dallas Value-Added Accountability System. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 81-99). Thousand Oaks, CA: Corwin Press, Inc.

Willms, D. J., & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement, 26*(3), 209–232.