

Discussion of Measurement Issues Associated with Value-Added Methods

**Michael J. Kolen
The University of Iowa**

Presented at a Workshop held by the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, Washington, DC: November 13-14, 2008.

Reckase and Linn Papers

- Provide important insights into measurement issues associated with value-added models
- Authors mainly agree, but differ in emphasis
- I agree with the perspectives provided
- I will emphasize a few issues that I think deserve further attention

Measurement Issues

- My comments are organized around the following measurement issues:
 - Measurement error
 - Test content
 - Score scales
 - Vertical scales
 - Validation of school and teacher effects

Types of Models

- Two general classes of models are affected differently by measurement concerns
- In *residual score models* (regression models), focus is on differences between predicted and observed scores (e.g., Sanders' VAM models)
- In *gain score models* focus is on differences between scores in current year and previous years (e.g., hierarchical growth models that focus on estimating growth trajectories)

Measurement Error

- Can distort regressions in residual score models (Reckase)
- Can lead to instability in teacher and school effects that have been documented in research (Linn)
- Can lead to differences in scores being unreliable for gain score models (Linn)
- Equating error can lead to instability in estimating teacher and school effects (NAEP Reading Anomaly; Waltman's work in Iowa)
- Often, measurement error is much larger at high and low scores and smaller at middle scores; might have a substantial effect on teacher and school effect estimates

Test Content

- Estimated teacher effects have been shown to “depend on skills that are measured by achievement tests” (Linn, p. 11)
- Test content likely has a substantial influence on school and teacher effects estimated from both gain and residual score models

Score Scales

- For all models an assumption of an equal interval scale is needed (Reckase and Linn)
- The claim that IRT scales are equal interval “is controversial and not easily verified” (Linn, p. 10); same could be said about any scale used with educational tests

Vertical Scale

- Is not necessary for residual score models but is required for gain score models (Reckase and Linn)
- Tests at different grade levels are purposefully built to differ in content and difficulty (Reckase and Linn); such differences likely influence estimates of teacher and school effects
- Briggs et al. (2008) found choice of vertical scale affects classification of schools into effectiveness categories
- Tong and Kolen (2007) found properties of vertical scales depend on how the data were collected and the statistical methods used
- Additional research should examine the influence of vertical scaling methods on school and teacher effects

Validation

- Teacher and school effects based on test scores can be viewed as measures of achievement at the aggregate level
- Viewed in this way, teacher and school effects should be subject to test validation efforts like those used with scores for individuals
- Current validation work focuses on procedural evidence such as reasonableness of statistical assumptions
- Other types of evidence should be gathered

Validation-Related Questions

- Are schools with higher school effect indices more effective than schools with lower school effect indices?
- What are the characteristics of teachers with higher and lower teacher effect indices?
- Are teacher effect indices sensitive to efforts made to educate teachers to be better teachers?
- How can the teacher and school effects be communicated in ways that lead to improved education?
- The results from which types of models are most easily translated into actions that can improve teachers, schools, and the general education of children?

Conclusion

- Measurement issues likely have a substantial affect on estimates of school and teacher effects
- The issues discussed in this session raise questions about the usefulness of estimates of school and teacher effects from the models considered
- Are estimated teacher and school effects due primarily to idiosyncrasies of the statistical methods, measurement error, the particular tests examined, and the scales used?
- Or, are estimated teacher and school effects due, at least in part, to educationally relevant factors?
- These questions should be answered clearly before these models are used to make important educational decisions