

Measurement Issues Associated with Value-Added Methods

Robert L. Linn

CRESST, University of Colorado at Boulder

Paper presented at a workshop held by the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability sponsored by the National Research Council and the National Academy of Education, Washington, DC: November 13-14, 2008

Value-Added Methods

Goal: To disentangle from the many factors that contribute to student achievement the effects that can be uniquely attributed to teachers, schools or educational programs

Promise of realizing goal explains appeal of VAM for teacher and school improvement, program evaluation, and teacher and school accountability

VAM Goal

- Realization of the goal of uniquely identifying teacher or school effects requires the justification of a causal inference
- Justification of causal inferences with non-random assignment a non-trivial undertaking that rest on a variety of assumptions
- My focus is more limited – measurement issues

Overview of Measurement Issues Addressed

- Individual and Aggregate Measurement Error
- Sensitivity to Instruction
- Growth Measures
- Scaling

Individual and Aggregate Measurement Error

- Measurement error at the individual student level is a familiar issue
 - Measurement error at the individual level is higher for difference scores used to estimate gains than for either of the scores used to compute the difference score
- Although measurement errors for individuals tend to average out for means, group level measurement error also can occur for mean scores at aggregate level of teacher or school

Individual and Aggregate Measurement Error (continued)

- Reliability of group averages may be higher or lower than reliability of individual scores (Zunbo & Forer, in press).
- Brennan, Yin, & Kane (2003) found g coefficients for relative difference scores for a single cohort of students ranged from a low of .25 for students per district to a high of .46 for 80 students per district

Individual and Aggregate Measurement Error (continued)

- Aggregate level measurement error deserves greater attention than it usually receives
- Ballou (2005) – only 40% of the math teachers with estimated teacher effects in the bottom quartile in 1998 were also in the bottom quartile in 1999 while “nearly a quarter of those in the top quartile in 1998 dropped below the median in 1999

Individual and Aggregate Measurement Error (continued)

- McCaffrey, Sass, & Lockwood (2008) found relatively low correlations (.05 to .35) between estimated teacher effects for successive years for both elementary and middle school teachers.
- Year-to-year instability due, in part, to measurement error and, in part, to real changes in teacher effectiveness
- Stability seems adequate of low-stakes uses of estimated teacher effects, but not for high-stakes uses.

Sensitivity to Instruction

- Results of value-added analyses are sensitive to the choice of tests used
- Lockwood, McCaffrey, Hamilton, Stecher, Le, & Martinez (2007) conducted separate value-added analyses for Procedures and Problem Solving subscales.
 - Correlations of teacher effect estimates were generally low
 - “... teacher performance may depend on the skills measured by the achievement test” (Lockwood, et al., 2007, p. 56).

Growth Measures

- Familiar dictum: “When measuring change, do not change the measure” (Beaton, 1990, p. 10).
- Value-added methods necessarily depend on the use of different tests at different grade levels
- Comparability of tests at different grade levels can be an issue

Growth Measures (continued)

- Vertical scales
 - “... connections between some levels are stronger than others, and sometimes the links between levels are too loose to maintain a sturdy connection between levels (Yen, 2007, p. 275).
 - “For mathematics, for example, tests at the 3rd grade measure predominantly arithmetic skills. By the 8th grade the test shifts to problem solving, pre-algebra, and algebra” (Reckase, 2004, p. 118)

Growth Measures (continued)

- Vertical scales
 - Changes in weights given to different constructs at different grades in vertically scaled tests undermines the validity of estimates of effects in value-added analyses (Martineau, 2006)
 - Results of value-added analyses are sensitive to the ways in which vertical scales are constructed (Briggs, Weeks, & Wiley, 2008)

Scaling

- Most value-added methods depend on the strong assumption that the test used have equal interval scales (Ballou, 2008; Reardon & Raudenbush, 2008)
- Although it is often claimed that IRT models used to scale tests result in equal interval scales, there are many reasons to question this claim

Scaling (continued)

- Ballou (2008) has suggested an ordinal approach to value-added analyses
- An ordinal approach may give up too much even if the scale does not satisfy the equal interval assumption exactly
- Alternative is to test the sensitivity of value-added results to monotonic transformations of test scores (Reardon & Raudenbush, 2008).

Conclusions

- Measurement Error
 - Low-stakes uses justifiable, but errors are too large for high-stakes uses for teachers
- Sensitivity to Instruction
 - Need to recognize that effect estimates depend on choice of test
- Growth Measures
 - Vertical scales, while not essential to value-added methods, raise a number of issues
- Scaling
 - Since equal interval assumption is likely to be violated, it seems desirable to routinely compare results of value-added analyses using different monotonic transformations of the test scale scores