Measurement Issues Associated with Value-Added Methods

Article · January 2008

CITATIONS	READS
0	47

1 author:

Robert L. Linn
University of Colorado Boulder
213 PUBLICATIONS **9,283** CITATIONS

Measurement Issues Associated with Value-Added Methods

Robert L. Linn

Center for Research on Evaluation, Standards, and Student Testing

University of Colorado at Boulder

Educational Accountability sponsored by the National Research Council and the National Academy of Education, Washington, DC: November 13-14, 2008.

Measurement Issues Associated with Value-Added Methods

Value-added methods have attracted considerable attention in the past few years from both policy makers and researchers. The goal of value-added methods is to be able to disentangle from the myriad factors that contribute to student achievement the effects that can be uniquely attributed to teachers, schools, or educational programs. Value-added methods employ a variety of sophisticated statistical techniques to accomplish the goal of identifying the effects that can be attributed to teachers, schools, or programs.

The promise of identifying unique teacher, school, or program effects obviously has great appeal to those interested in teacher and school improvement, teacher and school accountability, or program evaluation. Thus, the popularity of value-added models is not at all surprising. As with any effort to isolate causal effects from observational data where random assignment is not feasible, however, there are reasons to question the ability of value-added methods to achieve the goal of determining the value added by a particular teacher, school, or educational program.

Although the question of whether or not the applications of value-added methods justify causal claims is critically important, that question is beyond the scope of this paper. It is worth noting, however that several authors have challenged the causal claims made by value-added methods (e.g., Braun, 2005; Raudenbush, 2004; Reardon & Raudenbush, 2008; Rubin, Stuart, & Zanutto, 2004). The focus of this paper is more narrowly aimed at measurement issues in the application of value-added methods.

Individual and Aggregate Measurement Error

Individual student scores on achievement tests are the basic data that are used by value-added methods. Achievement tests are fallible measures. That is, observed student scores include some amount of measurement error. If an alternate form of the test were administered shortly after the first administration of the test, the scores on the two forms of the test would differ somewhat for the vast majority of students. Those differences are due to measurement error.

The magnitude of the measurement error can be estimated and used to provide a band of uncertainty for each student's test score. When tests are administered at two or more points in time, e.g., the spring of 3rd grade and the spring of 4th grade, the scores of each administration have a degree of uncertainty due to measurement error. When change or growth scores are computed by subtracting the score at time one from the score at time two, the difference scores also include measurement error. The measurement error for the difference scores is larger than the measurement error for either of its parts. Thus, individual student growth measures have more uncertainty due to measurement error than the measures of achievement at either point in time.

Although it is important to recognize the uncertainty due to measurement error at the individual student level, value-added methods focus on aggregate results, e.g., average results for students linked to a given teacher, a school, or an educational program. Consequently, the magnitude of the

measurement error associated with group means is relevant to an evaluation of results of value-added analytical results. If errors of measurement at the aggregate level were only due to the average of the individual student errors then the errors at the aggregate level would be quite small since the plus and minus errors for individual students would tend to cancel each other. Errors of measurement for group means, however, are not simply the average of individual student errors of measurement. As Zumbo and Forer (in press) noted, the reliability of group average scores may be higher or lower than the reliability of the individual student scores that are used to compute the group average.

Zumbo and Forer based their conclusion on the work of Kane and Brennan and their colleagues that has used generalizability theory to develop estimates of variance components and generalizability coefficients for group means during the past 30 years (see, for example, Kane & Brennan, 1977 for an early reference and Haertel, 2006, pages 95-97 for a recent discussion of the applicability of generalizability theory for group means and multilevel data). An example of the use of generalizability theory to estimate the magnitude of different sources of error and generalizability coefficients for group means that is particularly relevant for the topic of this paper is a study conducted by Brennan, Yin, & Kane (2003) that used data from the Iowa Tests of Basic Skills to investigate the dependability of district level differences in mean scores from one year to the next.

The dependability of the difference in means scores is relevant to value-added methods estimate the contributions of teachers, schools, or programs to changes in means from one year to the next. Brennan, Yin, and Kane (2003) found that the generalizability coefficients for relative difference scores and a single cohort of students ranged from a low of .254 with 20 students per district to .464 with 80 students per district. By conventional standards those coefficients are relatively low, especially when it is considered in the context of the number of students per teacher or per school at a given grade level with complete data. The level of generalizability improved somewhat when Brennan, Yin, and Kane 92003) used multivariate generalizability theory, but the degree of uncertainty for the mean difference scores was still substantial, suggesting that it is important to consider aggregate level errors in interpreting the results of value-added analyses.

Ballou (2005) reported the stability of the quartiles in which the estimated teacher effects were located in 1998 and 1999 for elementary and middle school teachers in a moderately large district in Tennessee. He found that 40% of the mathematics teachers whose estimated teacher effects ranked in the bottom quartile in 1998 were also in the bottom quartile in 1999, however 30% of those teachers ranked above the median in 1999. Although the stability was somewhat better for teachers who ranked in the top quartile in 1998, "nearly a quarter of those who were in the top quartile in 1998 dropped below the median the following year" (Ballou, 2005, p. 288). The observed instability is due, in part, to measurement errors at the individual student and aggregate

levels and, in part, to real changes in teacher effectiveness. Nonetheless, the level of instability is a potential issue for using the estimated teacher effects in a given year for purposes of teacher accountability or making decisions about teacher pay. On the other hand, the degree of stability appears to be adequate for the types of low-stakes uses for teacher improvement that are actually made of the value-added results in Tennessee.

McCaffrey, Sass, and Lockwood (2008) recently investigated the stability of teacher effect estimates based on value-added analyses from one year and cohort of students to the next (e.g., the estimated teacher effect estimates in 2000-01 compared to those in 2001-02) for elementary- and middle-school teachers in four counties in Florida. They computed 12 correlations (4 counties by 3 pairs of years) for elementary-school teachers and 16 correlations (4 counties by 4 pairs of years) for middle-school teachers. For elementary-school teachers the 12 correlations between estimates in consecutive years ranged from .09 to .34 with a median of .25. For middle-school teachers the 16 correlations ranged from .05 to .35 with a median of .205). Thus the year-to-year stability of estimated teacher effects might be characterized as being low to moderate. In practical terms the degree of stability is sufficient to justify low-stakes uses of the results for teacher estimated effects for purposes of teacher improvement but inadequate for high-stakes accountability purposes.

Sensitivity to Instruction

Interpretations of the results of applications of value-added methods depend on what the tests used in the analyses measure. The sensitivity of the

results to what is measured was demonstrated by Lockwood, McCaffrey, Hamilton, Stecher, Le, and Martinez (2007). Lockwood and his colleagues compared the results of value-added analytical results for a large school district using two different subscales. Separate analyses were conducted with the Procedures and the Problem Solving subscales of the Stanford mathematics assessments for grades 6, 7, and 8. They used a wide range of value-added models ranging from simple gain scores to models that used a variety of control variables. A total of 20 different value-added analytical approaches were applied to data regarding gains from year 1 to year 2 and gains from year 2 to year 3.

The estimated teacher effects for the two different measures had generally low correlations for both data sets regardless of which value-added method was used to calculate the estimated effects. The authors concluded that their "results provide a clear example that caution is needed when interpreting estimated teacher effects because there is the potential for teacher performance to depend on the skills that are measured by the achievement tests" (Lockwood, et al., 2007, p. 56).

Growth Measures

The simplest measure of growth is obtained by subtracting the score at time one from the score at time two. It only makes sense to subtract one score from another, however, when the two measures are exchangeable. A familiar dictum "When measuring change, do not change the measure" (Beaton, 1990, p. 10) came from extensive analyses of National Assessment of Educational

Progress (NAEP) reading anomaly that occurred in the 1980s. The NAEP

reading anomaly led to extensive investigations from which it was concluded

that unexpectedly large declines in reading achievement at ages 9 and 17 were

attributable to the combined effects of a number of seemingly small changes in

the reading assessments.

    Although the dictum from the investigations of the NAEP reading

anomaly was in the context of comparing aggregate results for successive

cohorts of students rather than in the context of value-added analyses using

longitudinal data, the idea that measures of change make the most sense when

the measures are the same or equivalent at both points in time in applicable in

both contexts. Of course, there are many reasons to want to change the

measures In the context of NAEP there is always pressure to make the new

measure better aligned with current instruction and to improve the

measurement characteristics of the assessment. In the context of applications of

value-added methods, the tests that are used at the end of one grade are not

suitable for use at the end of the next grade because students at the higher

grade have been learning content appropriate for the higher grade and the test

needs to reflect that content. But there must be some degree of comparability

of scores on the fifth grade test with those on the fourth grade test, if gains from

fourth to fifth grade are to be computed and used in value-added analyses.

    One approach to constructing the scores from tests used at different

grades so they are comparable is to create a vertical scale that spans several

grade levels. "A *vertical scale* (also referred to as a *developmental scale*) is an

extended score scale that spans a series of grades … and allows the estimation of student growth along a continuum" (Young, 2006, p. 469). Tests that are constructed for use at different grade levels are not strictly equivalent in the sense that two forms of the SAT might be considered to be. The tests used at different grade levels obviously differ in difficulty and content coverage by design.

Although we often act as if educational achievement tests were unidimensional and use item-response theory that assumes unidimensionality to construct scales, educational achievement tests are multidimensional. Paralleling changes in the curriculum from grade to grade the relative emphasis on different dimensions changes across grade levels. "Thus, the scale bends or curves through space. Connections between some levels are stronger … than others, and sometimes links between levels are too loose to maintain a sturdy connection between the test levels" (Yen, 2007, p. 275).

In relation to the issue of multidimensionality, Reckase (2004) has noted that the levels of vertically-scaled achievement tests differ in the mix of constructs that they measure. "For mathematics, for example, tests at the 3rd grade measure predominately arithmetic skills. By 8th grade, the test shifts to problem solving, pre-algebra and algebra skills" (Reckase, 2004, p. 118). Thus, vertically scaled tests cannot be equated. The linkage between tests designed for use at different grades is much weaker than equating (see, for example, Linn, 1993, Mislevy, 1992).

Martineau (2006) studied the effects of using tests that shift constructs across grades on the results of value-added analyses. He found that shifting construct mix across grades can have serious consequences when vertical scales are used in value-added analyses. The changes in the weights given to different constructs in the vertically scaled tests undermine the validity of the estimates of effects in value-added analyses. Based on his analyses, Martineau (2006) concluded that "there are no vertical scales that can be validly used in high-stakes analyses for estimating value-added to student growth in either grade-specific or student-tailored construct mixes – the two most desirable interpretations of value added to student growth" (p. 57).

Briggs, Weeks, and Wiley (2008) constructed eight different vertical scales that differed with respect to the item response theory (IRT) model used, the method used to estimate student scale scores, and the IRT calibration method used to place items from the different grades on the vertical scale. Although the estimated school effects from the value-added analyses were highly correlated for the eight vertical scales, the estimated school effects differed for the different scales. Briggs, et al. (2008) "found that the numbers of schools that could be reliably classified as effective, average, or ineffective was somewhat sensitive to the choice of the underlying vertical scale" (p.26).

Although there are many reasons to think that the vertical scaling of tests used at different grade levels does not result in strictly exchangeable scores on the tests at different levels, vertical scaling has yielded results that provide a rough sense of the magnitude of the changes in student achievement from one

grade to the next. They can be used to provide a general sense of the amount of growth that takes place from grade to grade. Vertical scales can be useful additions to value-added methods, but it is important to note that they are not a necessary part of the more sophisticated value-added methods. Analyses used for the value-added method made popular by the Tennessee value-added assessment system (TVASS) (Sanders and Horn, 1998), for example, does not require a vertical scale.

Scaling

The most used value-added methods depend on the strong assumption that the tests used in the analyses are equal-interval scales (Ballou (2008; Reardon and Raudenbush, 2008). That is, a 10 point increase from 30 to 40 must represent the same gain as a 10 point gain from 60 to 70 or any other region of the scale. It is clear that there are a number of scales that are used to report test scores, such as percentile ranks, or grade-equivalent scores are not equal-interval scales. Scales developed using IRT are often claimed to be equal interval, but this claim is controversial and cannot be easily verified.

Ballou (2008) has suggested that those concerned about violations of the equal-interval assumption might replace the usual value-added analyses with and ordinal value-added analysis. Rather than comparing gains in mean achievement for a teacher or school to the average for teachers or schools

comparisons could be made in terms of the proportion of students in a school or associated with a teacher that outperform the average teacher or school.

The use of an ordinal method as suggested by Ballou (2008) provides a viable option, but may be seen as giving up too much by those that believe that even if the scale used does not satisfy the equal interval assumption exactly, that it is still meaningful to compute means. As Reardon and Raudenbush (2008) suggest, however, "it would be useful to know the extent to which inferences from value-added models are sensitive to monotonic transformations of test scores" (p. 22) in situations where there are questions about the assumption that the tests are reported on an equal-interval scale.

Discussion

Measurement issues have considerable relevance to the evaluation of value-added methods. Measurement error occurs at both the individual student level and the aggregate level, and errors at both these levels can undermine the trustworthiness of the results of value-added methods. The magnitude of these errors is not so large that they swamp the signal in value-added analyses. The results can generally be considered to be trustworthy enough for low-stakes uses such as providing information useful for purposes of instructional improvement. The magnitude of individual and aggregate level measurement error is large enough to call into question more high-stakes uses for purposes of teacher accountability, however.

It is clear that all achievement tests do not measure the same thing and it is equally clear that the results to value-added analyses depend on the test that is used. Tests that measure different constructs are likely to yield different results. Thus, it is important that the selection of the tests to be used be made with careful attention to the content covered and the sensitivity of the tests to instructional interventions.

Although vertical scales are not essential to the use of value-added methods, they do aide in the interpretation of simple year-to-year gains in student achievement. If a vertical scale is to be used, however, it is important to recognize that the results of value-added analyses may be sensitive to the particular characteristics of the vertical scale that is used.

Commonly used value-added methods are based on the assumption that the test scores are on an equal-interval scale. The equal-interval assumption precludes the use of some types of scales (e.g. percentile rank or grade-equivalent scales) that obviously lack the equal-interval property and favors the choice of a IRT scales that purport to yield equal interval scales. Since the claim that any test scale is actual an equal interval scale is controversial, it seems prudent to investigate the sensitivity of the results of value-added analyses to violations of the equal-interval assumption, by routinely comparing the results of analyses using different monotonic transformations of the scale scores.

References

Ballou, D. (2005). Value-added assessment: Lessons from Tennessee. In R. W. Lissitz (Ed.), *Value-added models in education: Theory and applications* (pp. 272-297). Maple Grove, MN, JAM Press.

Beaton, A. E. (1990). Introduction. In A. E. Beaton & R. Zwick (Eds.), *The effect of changes in the National Assessment: Disentangling the NAEP 1985*86 reading anomaly. Revised.* ERIC ED 322 206.

Braun, H. (2005). Value-added modeling: What does due diligence require? In R. W. Lissitz (Ed.), *Value-added models in education: Theory and applications* (pp. 19-39). Maple Grove, MN, JAM Press.

Brennan, R. L., Yin, P., & Kane, M. T. (2003). Methodology for examining the reliability of group difference scores. *Journa l of Educational Measurement*, *40*, 207-230.

Briggs, D. C., Weeks, J. P., & Wiley, E. (2008). *The sensitivity of value-added modeling to the creation of a vertical score scale.* Paper presented the National Conference on Value-Added Modeling, Madison, WI, University of Wisconsin at Madison, April 22-24.

Haertel, E. H. (2006) Reliability in R. L., Brennan (Ed.), *Educational Measurement, (4th ed)*, pp. 65-110. Westport, CT: American Council on Education/Praeger.

Kane, M. T. & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research, 47*, 267-292.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83-102.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., &
Martinez, F. (2007). *Journal of Educational Measurement*, *44*, 47-67.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal.
vertically scaled student achievement data for growth-based, value-added
accountability. *Journal of Educational and Behavioral Statistics*, *31*(1), 35-62.

McCaffrey, D. F., Sass, T. R., & Lockwood, J. R. (2008). *The intertemporal stability
of teacher effect estimates*. Paper presented the National Conference on
Value-Added Modeling, Madison, WI, University of Wisconsin at Madison,
April 22-24.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, and
prospects*. Princeton, NJ: Educational Testing Service.

Raudenbush, S. W. (2004b). What are value-added models estimating and
what does this imply for statistical practice? *Journal of Educational and
Behavioral Statistics*, *29* (1), 121-129.

Reardon, S. F. & Raudenbush, S. W. (2008). *Assumptions of value-added models
for estimating school effects.* Paper presented the National Conference on
Value-Added Modeling, Madison, WI, University of Wisconsin at Madison,
April 22-24.

Reckase, M. D. (2004). The real world is more complicated than we would like.
Journal of Educational and Behavioral Statistics, *29*, (1), 117-120.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view
of value-added assessment. *Journal of Educational and Behavioral Statistics*, *29*
(1), 103-116.

Sanders, W. & Horn, H. S. (1998). Research findings from the Tennessee value-added assessment system (TVASS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, *12*(3), 822-828.

Yen W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273-283). New York: Spinger.

Young, M. J. (2006). Vertical scales. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 469-485). Mahwah, NJ: Lawrence Erlbaum Associates.

Zumbo, B. D. & Forer, B. (in press). Testing and measurement from a multilevel view: Psychometrics and validation. In J. Bovaird, K, Geisinger, & C. Buckendahl (Eds.), *High stakes testing in education – Science and practice in K-12 settings [Festschrift to Barbara Plake]*. Washington, DC: American Psychological Association Press.