

Discussion of Measurement Issues Associated with Value-Added Methods¹

Michael J. Kolen
The University of Iowa

Draft, October 26, 2008

The papers by Mark Reckase (Reckase, 2008, November) and Robert Linn (Linn, 2008, November) provide many important insights into measurement issues associated with value-added models. The two authors have few, if any, areas of disagreement, although the papers differ in emphasis. I agree with most of the perspectives provided, and will emphasize a few issues that I think deserve further attention.

My comments are organized around the following measurement issues that are considered in these papers: Effects of measurement error, test content, score scales, and vertical scales. I also discuss the issue of the validation of indices of school and teacher effects.

In discussing measurement issues associated with value-added models, it is important to distinguish between two types of models. In *residual score models*, scores in the current year are predicted from scores in previous years (and possibly other variables), using regression models. Differences between current year observed scores and current year predicted scores (i.e. residuals) are the basis for estimates of student, school, and teacher effects. Value-added models such as those described by Sanders and Horn (1998) are of this type.

In *gain score models*, the difference between scores in the current year and previous years are used as the basis for estimates of student, school, and teacher effects. In more complex gain score models, growth trajectories are estimated, often within a hierarchical linear model framework.

The effects of various measurement considerations differ for these two types of models. For example, a vertical scale is required when using gain score models, whereas, as Linn (2008, November) points out, a vertical scale is not necessary for residual score models. In addition, models from each of these types can produce very different estimates of school, teacher, and student effects. For example, Li and Kolen (2008, March) found quite different orderings of schools based on school effects estimated by gain score and residual score models. Because the measurement considerations are so different for the two types of models and the results can differ substantially, I distinguish between these model types, where appropriate, in the following discussion.

Effects of Measurement Error

Both Reckase and Linn note that measurement error can affect estimates from gain score and residual score models. For residual score models, Reckase

¹ Paper prepared for the Workshop held by the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, Washington, DC: November 13-14, 2008.

points out that estimates of regressions obtained from scores that contain measurement error can differ from regressions for the variables measured without measurement error. Therefore, the residual score models in use do not account for measurement error, which could distort the results.

Linn notes that measurement error likely is contributing to the research finding that teacher effects are not very stable from year-to-year. Linn indicates that the lack of stability suggests that the teacher effects obtained from residual score models are not sufficiently stable for high-stakes use.

Regarding gain score models, Linn discusses the issue of the reliability of difference scores. He notes that at the student level, the reliability of difference scores is less than the reliability of scores on each test. He also makes the important point that the degree of uncertainty in mean difference scores at the teacher or school level can be substantial. Thus, measurement error can have a substantial influence on school and teacher effects using gain score models.

One issue that neither author discusses is the affect of equating error on school and teacher effect estimates. For reasons of test security, different test forms (sets of test questions) are used in different years. The statistical process of test equating (Holland & Dorans, 2006; Kolen & Brennan, 2004) is used to adjust score conversions so that scores on different test forms can be used interchangeably. Because the gain score and residual score models use data across different years, and likely different test forms, equating error can affect the stability of the estimated school and teacher effects. Because it can affect all scores earned on a particular test form, equating error can be particularly problematic for aggregated scores, such as school and teacher effects. The National Assessment of Educational Progress (NAEP) reading anomaly (Zwick, 1991) was, in part, due to equating error. In examining trends in performance in Iowa schools at the state level, Waltman (2008, June) found that it was important to assess trends using examinees who were assessed over time with the same test form; otherwise, equating error appeared to distort the trends. Thus, equating error might add substantial error to the estimates of teacher and school effects with both residual score and gain score models.

Another issue that neither author discusses is the well-known finding that the variability of conditional errors of measurement differs, often substantially, along the score scale. For scales used with residual score and gain score models, conditional measurement error variability typically is smaller at middle scores and larger at extreme scores. Such differences in measurement error variability along the score scale could lead to much greater error in estimates of effects for high- or low-scoring classrooms or schools than for middle-scoring classrooms or schools. To my knowledge, differences in the variability of conditional errors of measurement on school and teacher effects estimated by these models have yet to be studied.

Test Content

Linn discusses the paper by Lockwood et al. (2007) that showed that teacher effects estimated were quite different for Mathematics Procedures versus Mathematics Problem Solving subtests of the Stanford Achievement Test. According to Linn (2008, November, p. 11), these findings demonstrate that estimated teacher

effects “depend on the skills that are measured by achievement tests.” Linn (2008, November, p. 11) concludes that “tests that measure different constructs are likely to yield different results” when used with these models. Much more research is needed to better understand the influence of the test used on the results from these models. Test content likely has a substantial influence on the school and teacher effects from both gain score and residual score models.

Score Scales

Both authors stress that the use of residual score and gain score models assumes that test scores are reported on a scale that has equal interval properties. Linn (2008, p. 10) makes the important point that the claim that IRT scales are equal interval “is controversial and cannot easily be verified.” The same could be said for normalized scores or for any other score scales used with educational tests. Thus, it is crucial that the sensitivity of estimated teacher and school effects to the choice of scale be studied. In addition, the ordinal methods for residual score models that Linn mentioned should be evaluated.

Vertical Scaling

As Linn indicates, vertical scales are not a necessary component of residual score models. However, both Linn and Reckase point out that vertical scales are a requirement for gain score models. They suggest that vertically scaled achievement tests are designed to be different, both in difficulty and in assessed content. For tests designed in this way, the construct may shift over grades. Such issues likely have consequences for teacher and school effects estimated using gain score models that are, of necessity, based on vertically scaled tests.

Briggs, Weeks, and Wiley (2008, June) found that the choice of vertical scale affected the classification of schools into effectiveness categories based on school effects, although the school effects using different vertical scales were highly correlated. Tong and Kolen (2007) found that the properties of vertical scales including the amount of average year-to-year growth and within grade variability were quite sensitive to how the vertical scale was constructed, including the design for data collection and the scaling method used. More research is needed to understand how these sorts of factors might affect teacher and school effects from gain score models.

In addition to issues discussed in the papers, it is important to consider that vertical scaling results depend on how the test is constructed. In an ideal vertical scaling situation for a standards-based test, content standards for the test would be well defined within grade. In addition, content standards would be clearly articulated across grades. Such articulation would make it clear what content standards are assessed at each grade and what content standards are assessed at overlapping grades.

Such well-articulated content standards would be used in designing tests and in constructing the vertical scale. In one design for vertical scaling, common items are administered in adjacent grades. To a large extent, these common items drive the amount of observed grade-to-grade growth. More growth is observed if items

are chosen that clearly reflect what is taught in the adjacent grades. Less growth might be observed if the common items are not chosen carefully.

Unfortunately, most state standards-based testing programs do not have well-articulated content standards across grades. In part, this resulted because No-Child Left Behind (NCLB) standards-based tests were originally developed only at grades that were well separated (e.g., grades 4 and 8). Intermediate grades were added later, but in the process few, if any, states developed content standards that are well articulated across grades. Thus, most state standards-based tests are less than optimal for use with vertical scaling. Based on these considerations, it is important to consider test content, the articulation of test content across grade levels, and the design of vertically scaling studies when evaluating the use of gain score models.

Validation

Teacher and school effects based on test scores can be viewed as measures of achievement at the aggregate level. Viewed in this way, it can be argued that estimates of teacher and school effects should be subject to the same kind of validation efforts that are used with scores of individuals.

Most of the validation efforts that I am aware of are focused on what I would call procedural evidence (much like content validation evidence). Such efforts are based on assessing the reasonableness of the statistical assumptions made and on other statistical considerations in developing residual score and gain score models. Efforts have gone as far as assessing the stability (reliability) of teacher and school effect indices.

Developers of residual score and gain score methods should undertake a validation process. Some questions to address: Can it be demonstrated that schools with higher school effect indices are more effective than schools with lower school effect indices? What are the characteristics of teachers with higher teacher effects? What are the characteristics of teachers with lower teacher effects? If efforts are made to educate teachers to be more effective teachers, are teacher effect indices sensitive to these sorts of efforts? 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? Finding answers to these sorts of questions, and evaluating estimates of school and teacher effects in an overall validation framework and theory of teacher and school effectiveness is an important next step.

Conclusion

The papers in the present session raise questions about the usefulness of residual score and gain score models in estimating teacher and school effects. I think it is reasonable to ask the following questions: Are the teacher and school effects estimated by these models due primarily to a combination of the idiosyncrasies of the statistical methods chosen, measurement error, the particular tests examined, the scales used, and measures chosen? Or, are the teacher and school effects estimated by these models due, at least in part, to educationally

relevant factors? The existing literature does not seem to provide a clear answer to these questions. These questions must be answered clearly before these models are used to make important educational decisions.

References

- Briggs, D. C., Weeks, J. P., & Wiley, E. (2008, June). *The Sensitivity of value-added modeling to the creation of a vertical score scale*. Unpublished manuscript, University of Colorado, Boulder, Colorado.
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Westport, CT: American Council on Education and Praeger.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (Second ed.). New York: Springer.
- Li, D., & Kolen, M. J. (2008, March). *Models of individual growth for school accountability – An empirical comparison*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.
- Linn, R. L. (2008, November). *Measurement issues associated with value-added methods*. Paper presented at a Workshop held by the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, Washington, DC.
- Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, 44(1), 47-67.
- Reckase, M. D. (2008, November). *Measurement issues associated with value-added methods*. Paper presented at a Workshop held by the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, Washington, DC..
- 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.
- Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, 20(2), 227-253.
- Waltman, K. (2008, June). *Changes in Iowa student performance: 2006-07 vs. 2002-03. IARP Report #003*. Center for Evaluation and Assessment, The University of Iowa, Iowa City, IA. (Downloaded on 10-22-2008 from <http://www.education.uiowa.edu/cea/RecentFindings.htm>)
- Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice*, 10(3), 10-16.