

*The Inter-Temporal Variability of Teacher Effect Estimates (2009).*

McCaffrey, D.F., Sass, T.R., Lockwood, J.R., and Mihaly, K. NCPI Working Paper Series No. 2009-03. Nashville, TN.

School districts have begun using estimates of teachers' effects on student test scores (or "value-added") for diagnostic purposes and allotting monetary rewards. Such estimate must be precise enough to identify high- and low-performing teachers and predict future performance accurately. We study the inter-temporal variability in value-added measures for elementary and middle school mathematics teachers from five large Florida school districts. Consistent with worker productivity measures in other occupations, teacher estimates are moderately stable, with year-to-year correlations ranging from 0.2-0.5 for elementary school and 0.3-0.7 for middle school teachers. Thirty to 60 percent of the variation in measured teacher performance is due to sampling error from "noise" in student test scores. Persistent teacher effects account for about 50 percent (70 percent) of the variation not due to noise for elementary (middle) school teachers; other time-varying factors account for the remaining variance. However, observed teacher characteristics (e.g., experience, advanced degrees and professional development) explain little inter-temporal variation unrelated to sampling errors. Averaging estimates from two years greatly enhances the stability of the estimates and improves their predictive ability. We also explore the sensitivity of stability to the value-added model specification and the achievement test.

To view the entire report, please [click here](#).

---

*Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness: An Extended Analysis of the Rothstein Critique (2009b).* Koedel, C., and Betts, J. NCPI Working Paper Series No. 2009-01. Nashville, TN.

Value-added modeling continues to gain traction as a tool for measuring teacher performance. However, recent research (Rothstein, 2009, forthcoming) questions the validity of the value-added approach by showing that it does not mitigate student teacher sorting bias (its presumed primary benefit). Our study explores this critique in more detail. Although we find that estimated teacher effects from some value-added models are severely biased, we also show that a sufficiently complex value-added model that evaluated teachers over multiple years reduces the sorting-bias problem to statistical insignificance. One implication of our findings is that data from the first year or two of classroom teaching for novice teachers may be insufficient to make reliable judgments about quality. Overall, our results suggest that in some cases value-added modeling will continue to provide useful information about the effectiveness of educational inputs.

To view the entire report, please [click here](#).

---

*Value-Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation* (2009a). Koedel, C., and Betts, J. NCPI Working Paper Series No. 2008-21. Nashville, TN.

Value-added measures of teacher quality may be sensitive to the quantitative properties of the testing instruments upon which they are based. This paper focuses on the sensitivity of value-added to a particularly relevant testing-instrument property - test-score ceiling effects. Test-score ceilings are likely to be increasingly common in testing instruments across the country as education policy continues to emphasize proficiency-based reform. Encouragingly, we show that over a wide range of test-score ceiling severity, teachers' value-added estimates are only negligibly influenced by ceiling effects. However, as ceiling conditions approach those found in minimum-competency testing environments, value-added results are significantly altered.

To view the entire report, please [click here](#).

---

*An Empirical Analysis of Teacher Spillover Effects in Secondary School* (2009).  
Koedel, C. NCPI Working Paper Series No. 2008-25. Nashville, TN.

This paper examines whether educational production in secondary school involves joint production among teachers across subjects. In doing so, it also provides insights into the reliability of value-added modeling. Teacher value-added to reading test scores is estimated for four different teacher types: English, math, science, and social studies. While the initial results indicate that reading output is jointly produced by math and English teachers, post-estimation falsification tests debunk the math-teacher "effects" - that is, there is in fact no evidence of joint production in secondary school. The results offer a mixed review of the value-added methodology, suggesting that it may be useful in some contexts but not others.

To read this chapter, please [click here](#).

---

*Test Scaling and Value-Added Measurement* (2008).  
Ballou, D. NCPI Working Paper Series No. 2008-23. Nashville, TN.

As currently practiced, value-added assessment relies on a strong assumption about the scales used to measure student achievement, namely that these are interval scales, with equal-sized gains at all points on the scale representing the same increment of learning. Many of the metrics in which test results are expressed do not have this property (e.g., percentile ranks, normal curve equivalents). However, this property is claimed for the scales obtained when tests are scored according to Item Response Theory.

This claim requires that examinees and test items constitute, in the terminology of representational measurement theory, a conjoint structure. Unfortunately, it is difficult to confirm that this condition is met. In addition, end users of the data lack access to item-level data to test these assumptions

themselves. The best they can do is to check the plausibility of the resulting scales. On this count, IRT scales often do poorly. Reasonable rescalings have a substantial impact on students' measured growth.

Methods of ordinal data analysis can be employed instead, on the weaker assumption that IRT scales permit us to rank students. Instead of comparing mean achievement of a teacher's students to the students of a (hypothetical) average teacher, ordinal analysis asks what fraction of the former outperforms the latter. The feasibility of such methods for value-added analysis is demonstrated. It is seen that value-added estimates are sensitive to the choice of ordinal methods over conventional techniques.

Clearly, if IRT scores are an interval-scaled variable, ordinal methods throw away valuable information. Practitioners of value-added measurement should ask themselves, however, whether they are so confident of the metric properties of these scales that they are willing to attribute differences between regression-based estimates of value added and estimates based on ordinal analysis to the superiority of the former.

To view the entire report, please [click here](#).

---

*Exploring Student-Teacher Interactions in Longitudinal Data* (2008).

Lockwood, J., and McCaffrey, D. NCPI Working Paper Series No. 2008-24. Nashville, TN.

This paper develops a model for longitudinal student achievement data designed to estimate heterogeneity in teacher effects across students of different achievement levels. The model specifies interactions between teacher effects and students' predicted scores on a test, estimating both average effects of individual teachers and interaction terms indicating whether individual teachers are differentially effective with students of different predicted scores. Using various longitudinal data sources, the authors find evidence of these interactions that are of relatively consistent but modest magnitude across different contexts, accounting for about 10 percent of the total variation in teacher effects across all students. However, the amount that the interactions matter in practice depends on how different are the groups of students taught by different teachers. Using empirical estimates of the heterogeneity of students across teachers, they find that the interactions account for about three to four percent of total variation in teacher effects on different classes, with somewhat larger values in middle school mathematics. These findings suggest that ignoring these interactions is not likely to introduce appreciable bias in estimated teacher effects for most teachers in most settings. The results of this study should be of interest to policymakers concerned about the validity of value-added measurement teacher effect estimates.

To view the entire report, please [click here](#).

\*This paper was presented at the Wisconsin Center for Education Research's National Center on Value-Added Measurement in March 2008.

---

*The Inter-Temporal Stability of Teacher Effect Estimates (2008).*

McCaffrey, D., Sass, T., and Lockwood, J. NCPI Working Paper Series No. 2008-22. Nashville, TN.

School districts have begun using estimates of teachers' effects on student test scores (or "value added") for diagnostic purposes and allotting monetary rewards. Such estimate must be precise enough to identify high- and low-performing teachers and predict future performance accurately. We study the inter-temporal variability in value-added measures for elementary and middle school mathematics teachers from five large Florida school districts. Consistent with worker productivity measures in other occupations, teacher estimates are moderately stable, with year-to-year correlations ranging from 0.2-0.5 for elementary school and 0.3-0.7 for middle school teachers. Thirty to 60 percent of the variation in measured teacher performance is due to sampling error from "noise" in student test scores. Persistent teacher effects account for about 50 percent (70 percent) of the variation not due to noise for elementary (middle) school teachers; other time-varying factors account for the remaining variance. However, observed teacher characteristics (e.g., experience, advanced degrees and professional development) explain little inter-temporal variation unrelated to sampling errors. Averaging estimates from two years greatly enhances the stability of the estimates and improves their predictive ability. We also explore the sensitivity of stability to the value-added model specification and the achievement test.

To view the entire report, please [click here](#).

\*This paper was presented at the Wisconsin Center for Education Research's National Center on Value-Added Measurement in March 2008.

---

*Value-Added and Other Methods for Measuring School Performance: An Analysis of Performance Measurement Strategies in Teacher Incentives Fund Proposals (2008).* Meyer, R., and Christian, M. NCPI Working Paper Series No. 2008-17. Nashville, TN.

One of the central challenges of designing and implementing a performance pay program is developing an approach for determining which schools, teachers, and administrators have performed well enough to have earned a bonus. The U.S. Department of Education's Teacher Incentive Fund (TIF) program provides grantees substantial latitude to create incentive pay systems that fit their local needs. This paper reviews the methods proposed by TIF grantees for measuring the performance of schools, teachers, and administrators with respect to student achievement. One of the major objectives of this paper is to evaluate the different performance measurement approaches in terms of a specific statistical standard - a value-added model. To simplify this analysis, authors focus primarily on value-added models of grade-level performance. Most, if not all, of the papers' conclusions also apply to value-added models of classroom/teacher performance.

To view the entire report, please [click here](#).

---

*Do Teacher Effects Persist When Teachers Move to Schools With Different Socioeconomic Environments?* (2008). Sanders, W., Wright, S.P., Springer, M., and Langevin, W. NCPI Working Paper Series No. 2008-20. Nashville, TN.

A considerable amount of this attention has been focused on schools with large concentrations of economically disadvantaged and minority students because students in these schools, on average, do not have equal likelihood of being in a sequence of classrooms with the same level of instructional quality as students in other schools. Select states and districts have proposed and implemented financial incentives to recruit and retain highly-effective teachers in these high need schools. A key challenge against programs offering financial incentives to highly effective teachers to either move to, or remain in, a high need school is a lack of quantitative research on whether teachers produce comparable results when they move to a school with a different socioeconomic environment from their previous school. In an effort to provide systematic information on this important question, this paper examines the relationship between measures of teaching effectiveness before and after teachers change schools which service student populations with demographics different from the previous school.

To view the entire report, please [click here](#).

---

*Re-Examining the Role of Teacher Quality in the Educational Production Function* (2007). Koedel, C., and Betts, J. NCPI Working Paper Series No. 2007-03. Nashville, TN.

This study uses administrative data linking students and teachers at the classroom level to estimate teacher value-added to student test scores. We find that variation in teacher quality is an important contributor to student achievement - more important than has been implied by previous work. This result is attributable, at least in part, to the lack of a ceiling effect in the testing instrument used to measure teacher quality. We also show that teacher qualifications are almost entirely unable to predict value-added. Motivated by this result, we consider whether it is feasible to incorporate value-added into evaluation or merit pay programs.

To view the entire report, please [click here](#).