CALDER
NATIONAL CENTER for ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH
Tracking Every Student's Learning Every Year
Urban Institute
a program of research from The Urban Institute with:
Duke University
Stanford University
University of Florida
University of Missouri-Columbia
University of Texas at Dallas
University of Washington

BRIEF 4
NOVEMBER 2008

The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy

TIM R.SASS

There is little doubt that teacher quality is a key determinant of student achievement, but finding ways to identify and reward the best teachers has proved illusive. Traditionally, teacher compensation has been based on measurable characteristics of teachers, like years of experience or attainment of advanced degrees. However, recent research has found at best a weak link between such measures and student performance. These findings, combined with the increased availability of longitudinal data tracking teachers and their students over time, have lead to efforts to measure teacher quality and make teacher personnel decisions based on student test scores.

"Value-added" analysis seeks to measure teacher quality by estimating the impact of teachers on student achievement, holding constant other factors that affect current student performance, including contemporaneous student ability and effort, family inputs, peer influences and school quality as well as the prior contributions of family, peer, teacher and school inputs.

While much progress has been made, there is still disagreement over which statistical approaches are best and the extent to which they produce accurate or unbiased measures of a teacher's contribution to student learning.[1] Clearly, for any performance-based personnel system to provide the correct incentives and enhance teacher quality, it is necessary for value-added measures to accurately measure true teacher productivity. Avoiding systematic errors in evaluating teacher performance is not sufficient, however. If value-added measures of teacher quality are unbiased, yet highly variable, their efficacy in high-stakes personnel decisions will be limited. For example, some have proposed using value-added measures to determine which teachers are granted tenure and which are dismissed after an initial probationary period. If value-added measures vary over time, a tenure policy based on a short time frame could lead to the dismissal of many truly effective teachers and the retention of others who ultimately turn out to be relatively ineffective in boosting student achievement. Similarly, if variability in value-added measures over time leads to wide swings in who is rewarded, teachers will view merit-based pay plans as largely random, greatly reducing any incentive effects of pay-for-performance schemes.

In this brief I consider the stability of value-added measures, the factors that are associated with the degree of stability and the resulting implications for future research and policy. In a companion policy brief, Dan Goldhaber and Mike Hansen explore the long-run stability of value-added measures and the associated implications for tenure policies. In contrast, I focus on the stability of value-added measures over shorter time spans, across school districts, and over test instruments and consider the implications of teacher effect stability for the design and implementation of performance-based teacher compensation schemes.

HOW STABLE ARE ESTIMATED TEACHER EFFECTS OVER TIME?

There have been relatively few studies of the inter-temporal stability of teacher effects, but all the

extant analyses suggest that value-added measures of teacher performance are, at best, moderately stable over time. Two recent studies, Koedel and Betts (2007) and McCaffrey et al. (2008), estimate value-added models with teacher fixed effects and compare the quintile rankings of teacher value-added for adjacent years.[2] Koedel and Betts rank high school math teachers in San Diego by quintile while McCaffrey et al. perform similar rankings of elementary and middle school teachers in four large school districts in Florida. A summary of their results is presented in table 1. The findings are quite consistent across the two studies. About one quarter to one third of the teachers in the bottom and top quintiles stay in the same quintile from one year to the next while roughly 10 to 15 percent of teachers move all the way from the bottom quintile to the top and an equal proportion fall from the top quintile to the lowest quintile in the next year. Thus, for example, if bonuses were allotted to teachers ranked in the top 20 percent based solely on value added, at most a third would get bonuses two years in a row and about one in ten who received a bonus one year would be ranked in the bottom 20 percent of teachers the next year.

Another way of looking at the inter-temporal stability of teacher effects is to measure the correlation between an individual teacher's value-added score in two adjacent years. If measured teacher quality remained constant over time then the correlation would equal one. In contrast, if teacher effects in one year were completely unrelated to effects in the previous year the correlation would equal zero. Year-to-year correlations of estimated teacher effects for the four Florida counties studied by McCaffrey et al. are presented in table 2. The correlations in the first elementary school column are based on the same value-estimates used to construct the quintile tabulations presented in table 1. Across four pairs of years the correlations for both elementary and middle school math teachers generally fall in the range of 0.2 to 0.3; a modest correlation at best.

WHAT DETERMINES THE STABILITY OF TEACHER EFFECT ESTIMATES?

The observed variation in measured teacher performance over time is not necessarily a bad thing. While it is typically assumed that a "good" teacher will consistently outperform an average teacher over time, true teacher quality could vary

Table 1. Quintile Rankings of Estimated Math Teacher Effects in 2000/01 and 2001/2002: Percent of Teachers by Row

Ranking in 2000/01		Ranking in 2001/02				
		Bottom 20%	Second 20%	Third 20%	Fourth 20%	Top 20%
Bottom 20%	San Diego, CA	35	25	16	14	11
	Duval Co., FL	30	20	20	12	18
	Hillsborough Co., FL	29	23	20	17	11
	Orange Co., FL	34	23	23	10	10
	Palm Beach Co., FL	24	12	22	26	16
Top 20%	San Diego, CA	12	9	25	24	29
	Duval Co., FL	14	13	22	25	27
	Hillsborough Co., FL	10	13	18	29	31
	Orange Co., FL	7	19	17	26	31
	Palm Beach Co., FL	13	18	18	20	22

San Diego data are from Koedel and Betts (2007), table 9. They represent high school teachers and are based on an achievement model with student and school fixed effects. Data for Florida counties are for elementary school teachers with 10 or more students per year and are based on estimates of models with student fixed effects and student, peer and school time-varying controls. Data and estimation procedures are described in McCaffrey et al. (2008). Both the San Diego and Florida analyses use achievement gains from the Stanford Achievement Test.

Table 2. Year-to-Year Correlations in Estimated Teacher-by-Year Effects for Four Florida Counties

County	2000/01 and 2001/02	2001/02 and 2002/03	2002/03 and 2003/04	2003/04 and 2004/05
		Elementary		
Duval Co., FL	0.24	0.28	0.25	0.27
Hillsborough Co., FL	0.27	0.25	0.16	0.29
Orange Co., FL	0.31	0.34	0.30	0.36
Palm Beach Co., FL	0.16	0.08	0.21	0.21
		Middle		
Duval Co., FL	0.22	0.35	0.31	0.26
Hillsborough Co., FL	0.38	0.31	0.28	0.18
Orange Co., FL	0.36	0.28	0.31	0.24
Palm Beach Co., FL	0.24	0.32	0.28	0.26

Data are for elementary and middle school teachers with 10 or more students per year and are based on estimates of models with student fixed effects and student, peer and school time-varying controls. Data and estimation procedures are described in McCaffrey et al. (2008).

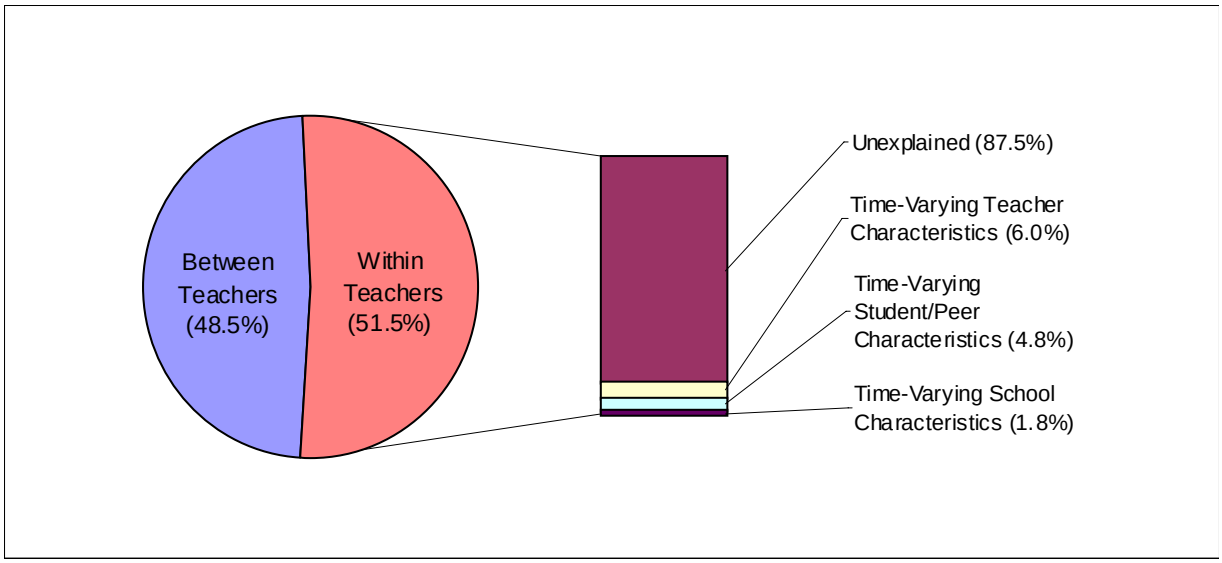
from year to year due to changes in teacher productivity associated with experience or acquisition of training from professional development or additional college coursework. If this is true then one would want teacher compensation to vary over time as well. However, if true teacher quality is indeed relatively invariant over time and measured productivity is simply due to "noise" (i.e. random error) in the estimated teacher effects, then compensation tied to value-added measures of teacher quality would fail to provide a strong link between teacher effort and reward.

McCaffrey et al. demonstrate that much of the variation in estimated teacher effects is in fact due to independent student-level variation in test performance over time, rather than changes in true teacher productivity. Within-student learning gains vary from year to year for reasons that cannot be explained by observable changes in student and family characteristics. Since value-added measures of teacher quality are essentially averages of student learning gains (accounting for *measured* student/peer and school characteristics) "unexplained" within-student variability in learning gains is transmitted into inter-temporal variation in teacher value-added. Indeed, McCaffrey et al. find that "dis-attenuated correlations," which attempt to estimate the correlation of teacher effects in the absence of student-level errors, lead to much higher correlations in year-to-year teacher effects, generally in the range of 0.5 to 0.8.

In order to better understand the sources of year-to-year variation in measured teacher effects, one can also decompose the variance in teacher-by-year value added into the components that can be explained by inter-temporal changes in the observed characteristics of students, teachers and schools. Figures 1A and 1B below illustrate the decomposition of the variance in estimated teacher-by-year effects for elementary and middle school math teachers, respectively.[3] The variation in measured teacher performance is broken down into variation across teachers and within teachers over time. The within-teacher variation is further divided into the proportions that can be explained by variation in students and their peers, teachers and schools over time and the remainder that is unexplained. The proportion of variation in estimated teacher effects attributable to within-teacher changes over time is much greater for elementary teachers (51.5 percent) than for middle school teachers (30.5 percent), suggesting greater inter-temporal stability in measured teacher quality at the elementary school level.

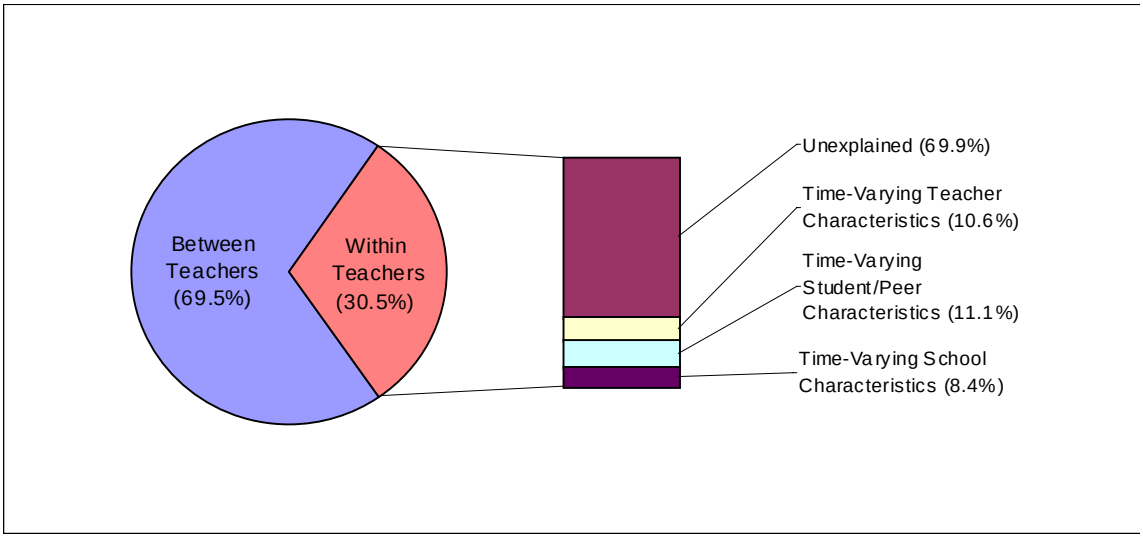
There are a number of possible explanations for this. Given the greater degree of tracking in middle school, variability in the unobserved traits of students may be greater in elementary school. Given that elementary teachers typically instruct the same group of students all day long, there may be stronger peer interactions and thus getting one or two students that generate negative spillovers may have a greater impact on the average performance of a teacher's students.[4] Further, there may be more

Figure 1A. Components of Estimated Elementary Math Teacher-by-Year Effects: Simple Average of Duval, Hillsborough, Orange, and Palm Beach Counties



Data are from McCaffrey et al. (2008), table 4.

Figure 1B. Components of Estimated Middle School Math Teacher-by-Year Effects: Simple Average of Duval, Hillsborough, Orange, and Palm Beach Counties



Data are from McCaffrey et al. (2008), table 4.

variation in unmeasured class-level supports, such as teacher aides or parent volunteers.

The proportion of within-teacher variance in estimated value-added that can be explained by observed time-varying factors is greater in middle school (30.1 percent) than elementary school (12.5 percent). However, even in the case of middle school, this still means that over two-thirds of the within-teacher variation in estimated value-added is due to inter-temporal variation in student, peer, teacher and school factors that affect student achievement, but which are not observed in the available data and hence cannot be accounted for when computing teacher value added. Put simply, there is considerable random error or "noise" in estimates of individual teacher value added which leads to variability in measured teacher effects over time.

ARE TEACHER EFFECT ESTIMATES SENSITIVE TO THE TEST INSTRUMENT?

Different states employ different achievement tests and often states change their test instrument over time. For example, both California and Texas have changed their statewide achievement tests in recent years. This begs the question of whether teacher rankings would differ much if a different achievement test is employed. Data from Florida over an opportunity to address the issue, as Florida administers both a high-stakes criterion-referenced exam, the "Sunshine State Standards" test (SSS) and a norm-referenced exam, the Stanford Achievement Test (NRT) to student in all grades 3–10.

Table 3 compares the rankings of elementary math teachers in Hillsborough County (Tampa area) based on the low-stakes NRT exam and the high-stakes SSS exam. There is less variability across tests at a point in time than with a given test over time. Nearly half (43 percent) of teachers ranked in the top quintile using the NRT are also ranked in the top quintile when computing student achievement using the SSS. Fully 70 percent of top-quintile teachers, as judged by student gains on the NRT, are ranked in the top two quintiles when using the SSS to compute teacher value added. Overall, the cross-exam correlation in estimated teacher effects is 0.48, much higher than the year-to-year correlation for the NRT exam of 0.27, reported in table 2.[5] Nonetheless, it is clear that different tests result in different teacher rankings.[6] This could be due to differences in the material being tested, differences in maximum measured achievement (i.e., "ceiling effects") or differential responses to the accountability pressures associated with the SSS exam that are not present with the NRT exam.

IMPLICATIONS FOR POLICY AND FUTURE RESEARCH

The findings that estimated teacher effects are relatively unstable over time and the instability is due largely to inter-temporal changes in unobservable student-level variables, should give one pause when considering the use of annual value-added measures in pay-for-performance programs. If rewards are tied to measures that fluctuate widely over time for reasons beyond a teacher's control, they are unlikely to produce significant incentives for teachers to increase their productivity. However, the relative instability of current value-added measures of student performance does not imply there can be no role for student outcomes in teacher compensation plans.

TABLE 3. CROSS-EXAM STABILITY OF ELEMENTARY MATH TEACHER EFFECT ESTIMATES BY QUINTILE (PERCENT OF TEACHERS BY ROW) – HILLSBOROUGH CO., FL, 2001/02 [CORRELATION = 0.48]

	Ranking Based on SSS				
Ranking based on NRT	Bottom 20%	Second 20%	Third 20%	Fourth 20%	Top 20%
Bottom 20%	43	26	14	11	5
Second 20%	26	25	23	17	10
Third 20%	15	21	21	24	18
Fourth 20%	11	19	25	20	24
Top 20%	4	9	17	27	43

NRT is the Stanford Achievement test, SSS is the "Sunshine State Standards" criterion-based exam tied to Florida's curriculum standards.

While current measures are quite unstable, it is possible statistical methods can be employed to reduce the impact of student-level "noise" in test scores on estimates of teacher value-added. The work of McCaffrey et al. suggests that this could significantly enhance the year-to-year correlation of estimated teacher effects. However, such procedures come at the cost of reducing the transparency of teacher quality measures. It is unlikely that teachers would buy into a system where the measure of teacher productivity is difficult to understand. Indeed current regression-based methods used to generate value-added estimates are rather opaque to most stakeholders. Adding additional layers of statistical complexity to adjust for "noise" in student test scores will likely make the whole system even less transparent. Ideally one would want to find a way to account for student-level test score variability that is easily understood by stakeholders. For example, basing teacher compensation on a multi-year average of value-added, rather than a single year, could help reduce the instability that occurs when using single-year measures. Averaging across years would also tend to smooth away true annual fluctuations in teacher performance, however.

Another option would be to adopt teacher compensation policies that rely on a mix of value-added performance measures and subjective evaluations by principals. Recent research by Harris and Sass (2007) and by Jacob and Lefgren (2008) indicates principals can effectively distinguish which teachers will have the largest impact on student achievement. Both studies find that principal evaluations are better predictors of teacher value-added than are teacher experience and educational attainment, the traditional measures used for teacher compensation. However, while Harris and Sass find that prior value-added measures and principal evaluations predict current student achievement gains equally well, Jacob and Lefgren find that past value-added estimates generally do a better job at predicting future student achievement than do subjective principal assessments. It is quite possible that a hybrid compensation system that included both value-added measures and subjective principal evaluations could outperform a system based on just a single measure. Determining the accuracy and stability of teacher ratings from such a hybrid system is an area that is ripe for future research.

Ultimately, one must evaluate the use of value-added measures in a teacher compensation system relative to other feasible alternatives. Despite their shortcomings, at least partial reliance on value-added measures when determining teacher compensation may prove to yield better student outcomes than the traditional compensation system which is based on teacher experience and educational attainment. Ongoing research on extant performance-pay systems for teachers should yield more definitive answers on the overall efficacy of using student performance as a component of teacher compensation.

REFERENCES

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 25:95–135.

Andrabi, Tahir, Jishnu Das, Asim Khwaja, and Tristan Zajonc. 2008. "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics." Unpublished manuscript.

Ballou, Dale. 2005. "Value-added assessment: Lessons from Tennessee." In R. Lissetz (Ed), *Value Added Models in Education: Theory and Applications*. Maple Grove, MN: JAM Press.

Burke, Mary A., and Tim R. Sass. 2008. "Classroom Peer Effects and Student Achievement." Working Paper 18. Washington, DC: Urban Institute, Center for Analysis of Longitudinal Data in Education Research.

Jacob, Brian A., and Lars Lefgren. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics* 26:101–36.

Koedel, Cory and Julian R. Betts. 2007. "Re-Examining the Role of Teacher Quality in the Educational Production Function." Working Paper #2007-03. Nashville, TN: National Center on Performance Initiatives.

McCaffrey, Daniel F., Tim R. Sass, and J.R. Lockwood. 2008. "The Intertemporal Stability of Teacher Effect Estimates." Unpublished manuscript.

Harris, Douglas N., and Tim R. Sass. 2007. "What Makes for a Good Teacher and Who Can Tell?" Unpublished manuscript.

———. Forthcoming. "The Effects of NBPTS-Certified Teachers on Student Achievement." *Journal of Policy Analysis and Management*.

Rivkin, Steven G. 2007. "Value-Added Analysis and Education Policy." Policy Brief 1. Washington, DC: The Urban Institute, Center for Analysis of Longitudinal Data in Education Research.

Rothstein, Jesse. 2008. "Teacher Quality in Educational Production." Working paper #14442. Cambridge, MA: National Bureau of Economic Research.

NOTES

[1] For a review of the value-added literature, see Rivkin (2007). Recent critiques of the literature include Andrabi et al. (2008) and Rothstein (2008).

[2] Two previous papers, Ballou (2005) and Aaronson et al. (2007), conduct similar analyses, breaking up teacher value-added rankings into quartiles, rather than quintiles. Despite differences in location, grade levels, and estimation methods, the findings are remarkably similar to one another and reinforce the findings presented above.

[3] Details of the computation of the variance decomposition are provided in McCaffrey et al. (2008).

[4] Burke and Sass (2008) find that classroom peer effects are generally larger in elementary school than in middle school.

[5] McCaffrey et al. also examine the impact of changing tests on the inter-temporal variability in teacher effect estimates. The observed variation in measured teacher performance in some cases changes significantly across tests. However, there is no consistent pattern across tests and school districts.

[6] Harris and Sass (forthcoming) find some differences in the estimated effectiveness of National Board certified teachers, depending on which test instrument is used to measure teacher quality.

ABOUT THE AUTHOR

Tim R. Sass is Professor of Economics at Florida State University specializing in applied microeconomics, industrial organization, public choice and the economics of education. Dr. Sass is a member of the CALDER Florida team. His work covers an array of education policy issues including charter schools, teacher quality, special education, vouchers, peer effects and value-added methodology.

THE URBAN INSTITUTE
2100 M Street, N.W.
Washington, D.C. 20037

Phone: 202-833-7200
Fax: 202-467-5775
http://www.urban.org

National Center for Analysis of Longitudinal Data in Education Research

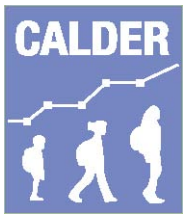
IN THIS ISSUE

The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy

For more information, call Public Affairs at 202-261-5709 or visit our web site, http://www.urban.org.

To order additional copies of this publication, call 202-261-5687 or 877-UIPRESS, or visit our online bookstore, http://www.uipress.org.

National Center for Analysis of Longitudinal Data in Education Research



This research is part of the activities of the National Center for the Analysis of Longitudinal Data in Education Research (CALDER). CALDER is supported by Grant R305A060018 to the Urban Institute from the Institute of Education Sciences, U.S. Department of Education. More information on CALDER is available at http://www.caldercenter.org.