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 3
NOVEMBER 2008

Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions

DAN GOLDHABER AND MICHAEL HANSEN
UNIVERSITY OF WASHINGTON BOTHELL, CENTER ON REINVENTING PUBLIC EDUCATION

USING TEACHER EFFECTS ESTIMATES FOR HIGH-STAKES PERSONNEL DECISIONS

Well over a decade into the standards movement, the idea of holding schools accountable for results is now being pushed to a logical, if controversial, end point: the implementation of policies aimed at holding individual teachers (not just schools) accountable for results. Some have called for reforms such as pay for performance or changes to teacher tenure to make it easier to reward and sanction teachers based on their classroom performance. The focus on teachers is more than just a logical extension of the standards movement. It is supported by two important findings from teacher quality research: teacher quality (as measured by teacher contributions toward student gains on tests) is the most important schooling factor when it comes to improving student achievement, and teacher quality is a highly variable commodity—some teachers are simply much better than others. These findings, coupled with a large body of research suggesting that typical characteristics used to determine employment and pay (such as experience and credentials) are not strongly correlated with effective teaching, are good reasons to move the policy discussion toward a focus on individual teacher performance.

The research on teacher effects (the terms "teacher effects" and "teacher job performance" are used interchangeably here) finds considerable variation in estimated job performance, suggesting there is great potential for improving education through *teacher workforce* accountability policies such as teacher tenure reforms, selective retention, salary incentives, and targeted professional development.[1] However, consensus about the right way to do this has been elusive. "Value-added" models (VAMs) seek to isolate the contribution that teachers make toward student achievement gains on tests and are increasingly being considered as a potential tool for evaluating teacher performance. The use of this metric is controversial, however—not only because of disagreement about whether student test scores ought to be used to judge teachers, but also because, even among those that support this use of student tests, there is no consensus on the *right statistical approach*, especially when it comes to estimating the impacts of individual teachers.[2] Moreover, relatively few studies have focused on the stability of estimated teacher effects, which have important implications for the formation of productive teacher policies.[3] For instance, if performance turns out to be an extremely stable characteristic, then measurement and accountability might best be used to weed out poor performers, as is suggested by Gordon and colleagues (2006). Alternatively, if actual performance (or our measures of it) tends to be an unstable characteristic—over time, across student types, or educational settings—then it may be necessary to rethink this direction for teacher-based accountability.

This research brief presents selected findings from work examining the stability of value-added model estimates of teacher effectiveness and their implication for tenure policies.[4] In related work, Steve Rivkin (2007) explores the use of value-added models for estimating true teacher

productivity, and Tim Sass (2008) assesses the stability of value-added measures of teacher productivity and their implications for pay-for-performance policies.

ESTIMATING TEACHER PERFORMANCE AND ITS STABILITY: DATA AND ANALYTIC APPROACH

In order to assess the stability of estimated teacher performance over time, it is necessary to have data that links students to their teachers and tracks them longitudinally. The data we utilize are collected by the North Carolina Department of Public Instruction (NCDPI) and managed by Duke University's North Carolina Education Research Data Center (NCERDC). These data, based on administrative records of all teachers and students in North Carolina, include information on student performance on standardized tests in math and reading (in grades 3 through 8) that are administered as part of the North Carolina accountability system.[5] We currently have data for teachers and students from school years 1995–1996 through 2005–2006.

Unfortunately, the North Carolina data do not include explicit ways to match students to their classroom teachers. They do, however, identify the proctor of each student's end-of-grade tests, and in elementary school the exam proctors are generally the teachers for the class. We utilize the listed proctor as our proxy for a student's classroom teacher, but we take several precautionary measures to reduce the possibility of inaccurate matches. First, we restrict our sample to those matches where the listed proctors have separate personnel file information and classroom assignments that are consistent with them teaching the specified grade and class for which they proctored the exam. We also restrict the data sample to self-contained, non-specialty classes, and because we wish to use data from classes that are most representative of typical classroom situations, we impose classroom restrictions limiting the size of the class to no less than 10 (to obtain a reasonable level of inference in our teacher effectiveness estimates) and no more than 29 (the maximum for elementary classrooms in North Carolina). Finally, we restrict our analyses to 5th grade teachers, because we are most confident in linking students to teachers in the elementary grades, and with 5th grade teachers we can use two years of testing history for students, minimizing the bias of scores as suggested in Rothstein (2008b).

These restrictions leave us a sample of 9,979 unique teachers and 29,003 unique teacher-year observations spanning 10 years (most teachers are observed more than once in the data). In this brief, we are focused primarily on the stability of teacher effects as they relate to tenure decisions. In North Carolina, state policy dictates that teachers receive tenure after teaching in the state's public schools for four consecutive years (Joyce 2000).[6] So, for the findings reported here, we restrict the sample further to teachers for whom we observe their first two years in the classroom (in North Carolina) and at least one year after they receive tenure. While we have nearly 10,000 unique 5th grade teachers for whom we can estimate teacher effectiveness, we observe only 1,363 unique novice teachers before 2003 (the last year for entering teachers for whom we could also observe post-tenure performance). Of these, only a small percentage stay in the teacher workforce long enough to observe post-tenure performance: 281 for whom we observe both post-tenure performance and performance estimates for their first two years of teaching, and 250 for whom we observe both post-tenure performance and performance estimates for their first three years of teaching.[7] Thus, our analysis sample represents a very select group of teachers, from which one should be cautious about drawing strong inferences about the teacher workforce in general.

Assessing the *stability* of estimated teacher effects, of course, requires the estimation of the teacher effects themselves. The research on VAMs suggests that teacher effect estimates are sensitive to model specification; so, while there is no universal standard for how these effects should be estimated, there are good reasons to believe that models using a full history of student test scores are likely to suffer from the least amount of bias. In fact, a recent paper by Kane and Staiger (2008) shows that this model specification produces teacher effects quite similar to those produced under conditions where teachers are randomly matched to their classrooms.[8] Thus, we utilize a value-added model of this form to calculate teacher effects.[9]

STABILITY OF TEACHER JOB PERFORMANCE ESTIMATES OVER TEACHING CAREERS

The existing research on the inter-temporal stability of teacher effects shows that the correlations of job performance estimates from one year to the next are thought to be "modest" (see Aaronson et al. 2007; Ballou 2005; Koedel and Betts 2007; and McCaffrey et al. 2008) and a nontrivial proportion of the inter-temporal variation in these effects appears to due to noise (see Sass 2008). In our recently released study (Goldhaber and Hansen 2008), we find similar results. Neglecting any time-varying variables in the model, we find that variation between teachers explains 52 percent of overall variance in reading and 63 percent in math.[10] And, the inclusion of time-varying teacher variables in the model has only a trivial effect on the within- versus between-teacher comparison, changing the measures by less than one tenth of one percent. In fact, only two teacher variables were found to be statistically significant predictors of within-teacher variation in effectiveness: a teacher's experience level and the number of discretionary absences. Thus, our decomposition suggests that changes in teacher quality within a teacher over time are, like teacher quality itself, almost entirely attributable to unobservable factors.

We find inter-temporal correlations of teacher effectiveness estimates of 0.30 in reading and 0.52 in math. Interestingly, these are not very different from estimates of job performance in sectors of the economy that consider them for high-stakes purposes (such as job retention and pay determination), where the year-to-year correlations of performance ranged from 0.26 to 0.55 (see Hoffmann et al. 1992, 1993).

There are good reasons to believe that *true* teacher effects might not be stable over a teacher's career. The consistency of job performance is likely greater for more-experienced employees, as familiarity with job tasks instills job behaviors that permit a smoother reaction to changes in job requirements (Deadrick and Madigan 1990). In the case of teachers, for instance, one commonly hears that "classroom management" is an essential tool learned early on in their careers. The acquisition of this or other skills appears to lead teachers to become more productive as they first gain experience (Clotfelter et al. 2006; Hanushek et al. 2005; Rockoff 2004), but also may lead to a narrowing of fluctuations in productivity. Also, one might imagine that teachers, as they settle into a particular setting, tend to adopt the practices of that setting (see Zevin 1974).

The above arguments suggest that one might expect a general convergence in teacher effectiveness as teachers become socialized into the norms of a school or district, or the profession. To investigate this possibility, we grouped teachers in our sample by experience level to see whether the variance in teacher effects differed by grouping. Although we do not report the details of this investigation here, in general we found no detectable change in the variance of teacher effects, and, in particular, no evidence of differences in variance pre- and post-tenure.

PRE- AND POST-TENURE TEACHER JOB PERFORMANCE RANKINGS

Whether early-career estimates of teacher effectiveness accurately predict later performance is of key interest to those who advocate allowing more individuals to initially enter the teaching profession and then being more selective about who is allowed to remain (Hanushek forthcoming; Gordon et al. 2006). Clearly an assumption underlying this proposal is that one can infer to a reasonable degree how well a teacher will perform over her career based on estimates of her early-career effectiveness.

To assess this issue, we group teachers into performance quintiles and explore the extent to which teachers tend to move from one quintile to another in pre- and post-tenure periods. As noted above, tenure in North Carolina is granted to teachers who teach full time in the state's public schools for at least four years. Thus, in principle it is possible to use all four years of teacher job performance information in considering whether to grant tenure, but in practice it is quite unlikely that four years of value-added calculations would be available. Moreover, in many states, tenure is granted after just three years of classroom teaching (and, in some states, even sooner). For these reasons, we calculate pre-tenure teacher effectiveness based on both a teacher's first two years and first three years in the classroom. We use this information to rank teachers, and then compare these rankings to their overall rankings for post-tenure job performance (a weighted average of all post-tenure observations).

Table 1 shows how teachers' pre-tenure reading performance rankings compare to their rankings post-tenure.[11] Consistent with other studies that report on adjacent year transitions, we find that teachers *tend* to stay in the same quintile category pre- and post-tenure, but there is also a fair amount of shifting around in the distribution. For instance, focus first on Panel A, which bases pre-tenure teacher effectiveness estimates on the first two years of teacher-student classroom data. Here we see that 32 percent of teachers in the lowest effectiveness quintile based on estimated pre-tenure job performance are also in the lowest effectiveness quintile based on estimated post-tenure job performance; and 46 percent of teachers in the highest quintile pre-tenure are in the highest quintile post-tenure. But there are a nontrivial proportion of teachers (11 percent) who are judged to be relatively ineffective (lowest quintile) based on the pre-tenure period who are then judged to be among the most effective (top quintile) in the post-tenure period.

Panel B of table 1 shows the same transitions when pre-tenure teacher job performance estimates are based on three years of data rather than two. While some small changes appear, we would argue that, on the whole, the addition of a third year of teacher job performance information changes the picture very little (though the correlations in the pre- and post-tenure teacher effects estimates are higher for the three year estimates).[12] While clearly a value-judgment, we would argue that this finding strongly suggests that two years of early-career job performance information is nearly as good as three for predicting later career performance (a point we return to below).

In panels A and B of table 2 we report the equivalent transition matrices for estimated teacher effectiveness in math. In contrast to our reading findings, we see significantly greater stability in pre- and post-tenure job performance, especially at the bottom of the distribution. For example, focusing on panel A, we see that over 44 percent of those teachers judged to be in the bottom quintile in pre-tenure math effectiveness were judged to be in the bottom quintile post-tenure (as opposed to 32 percent in reading). And, only 2 percent of teachers in the bottom quintile show up in the top quintile post-tenure (as opposed to around 11 percent in reading). However, similar to our reading findings, the distribution of teacher rankings changes only modestly when we move from the two-year pre-tenure job performance estimates (Panel A) to the three-year estimates (Panel B).[13] It is not surprising, given that these findings are based on multiple years of estimated teacher performance, that the stability of estimated math effectiveness pre- and post-tenure is considerably higher than what is

Table 1. Transition Matrices on Pre- vs. Post-tenure Estimated Teacher Job Performance: Reading

Pre-tenure quintile rank	Post-tenure Quintile Rank (percent)					Total teachers
	Bottom	Second	Third	Fourth	Top	
	Panel A. Using first two years of performance to predict post-tenure performance					
Bottom	32	23	19	16	11	57
Second	27	14	27	18	14	56
Third	21	23	30	18	7	56
Fourth	16	27	18	18	21	56
Top	5	13	5	30	46	56
Total teachers	57	56	56	56	56	281
	Panel B. Using first three years of performance to predict post-tenure performance					
Bottom	26	30	18	14	12	50
Second	28	14	38	12	8	50
Third	26	24	16	22	12	50
Fourth	12	18	22	24	24	50
Top	8	14	6	28	44	50
Total teachers	50	50	50	50	50	50

found for teacher math effects in other studies that focus on year-to-year changes in teacher quintile rankings (see, for instance, table 1 in Sass 2008).

TRADE-OFFS: USING PRE-TENURE JOB PERFORMANCE ESTIMATES TO DESELECT TEACHERS

The idea of teacher "deselection" has gained prominence in recent years. Economist Eric Hanushek, for instance, makes the case that school systems ought to be much more serious about using on-the-job teacher performance information to identify poorly performing teachers, and systematically "deselect" them from the workforce (forthcoming). His estimates suggest that students would greatly benefit if even a relatively small share of ineffective teachers were removed from the classroom and replaced with teachers who were of average performance. Using VAM estimates to help inform tenure decisions is an option that is often floated in policy discussions. Tenure is a natural point in a teaching career for making judgments about a teacher's future potential, given the greater difficulty of removing ineffective teachers once they are afforded the job protections that come from being tenured.

The extent to which receiving tenure is a rigorous screening process is hotly debated (for example, see http://edwize.org/tenure-travails). We cannot know from our data which teachers are deselected or tenured based on assessments of their performance.[14] However, we do know from looking at the entire North Carolina workforce in 2004 (as an example) that, of the 40,142 teachers in their 4th year, 34,120 (roughly 85 percent) moved into their 5th year having presumably been granted tenure.

Making the assumption that our estimates of teacher effectiveness are reasonable representations of true teacher job performance, the information provided in tables 1 and 2 shows the extent to which deselecting teachers based on pre-tenure performance would result in errors, in the sense that teachers who are poor performers in the pre-tenure period might be selected out of the workforce despite the fact that they turn out to be quite effective post-tenure.[15] The contrast in stability of estimated job performance between reading and math estimates is striking, but perhaps not terribly surprising since value-added models often predict less of the variation in student achievement in reading than in math. However, focusing on estimated teacher effects in reading and math separately raises a thorny issue in the case of using VAM in the context of tenure reform: at the elementary level, teachers are generally in self-

TABLE 2. TRANSITION MATRICES ON PRE- VS. POST-TENURE ESTIMATED TEACHER JOB PERFORMANCE: MATH

Pre-tenure quintile rank	Post-tenure Quintile Rank (percent)					Total teachers
	Bottom	Second	Third	Fourth	Top	
	Panel A. Using first two years of performance to predict post-tenure performance					
Bottom	44	25	14	16	2	57
Second	25	30	25	13	7	56
Third	14	14	30	18	23	56
Fourth	14	18	18	23	27	56
Top	4	13	13	30	41	56
Total teachers	57	56	56	56	56	281
	Panel B. Using first three years of performance to predict post-tenure performance					
Bottom	42	26	18	10	4	50
Second	36	28	20	12	4	50
Third	16	24	26	18	16	50
Fourth	4	14	20	28	34	50
Top	2	8	16	32	42	50
Total teachers	50	50	50	50	50	50

contained classrooms and responsible for students in all subject areas, so what happens if teachers are relatively more effective in teaching a particular subject area?

In Goldhaber and Hansen (2008), we explore the cross-subject correlation of estimated teacher effects and report findings in math and reading in the range of 0.50. But, it is worth investigating the potential implications of tenure policies that rely on estimates of teacher effects in both reading *and* math. Imagine, for instance, a more-stringent tenure policy that deselected teachers if they were in the lowest quintile in *either* reading or math effects. The first thing to note is that our calculation suggests that such a policy would result in approximately 30 percent of our sample not receiving tenure.

Figure 1 shows the implications of such a policy. In this case, we calculate that approximately half of those who would be excluded under such a policy have reading performance in the lower two quintiles post-tenure, and almost 58 percent have math performance in the lower two quintiles post-tenure. However, we also note that almost 30 percent of teachers in reading, and about a quarter in math, fall into one of the top two performance quintiles post-tenure.

Now instead consider a more relaxed tenure policy that deselected teachers only if they fall into the lowest quintile on pre-tenure estimates of *both*

FIGURE 1. DE-SELECTING TEACHERS WITH LOW PERFORMANCE ESTIMATES: EITHER SUBJECT

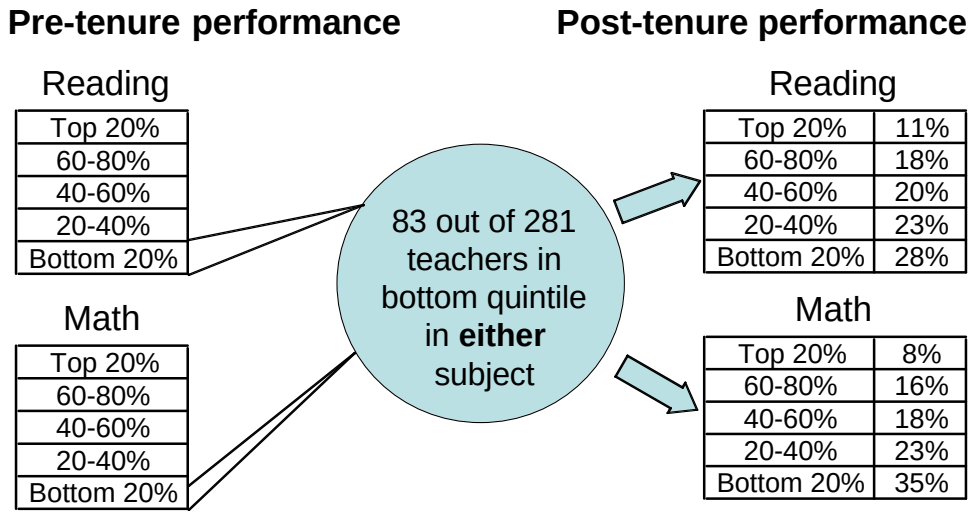
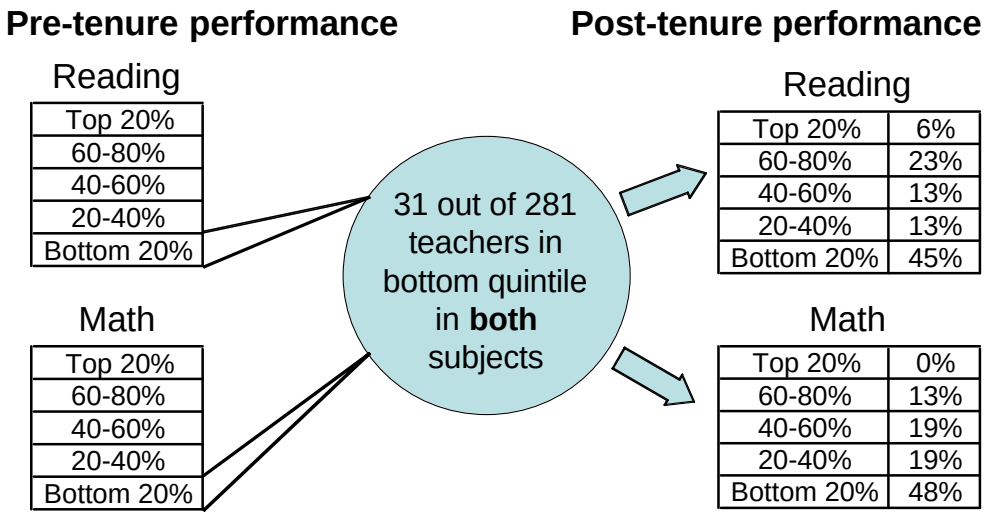


FIGURE 2. DE-SELECTING TEACHERS WITH LOW PERFORMANCE ESTIMATES: BOTH SUBJECTS



reading and math effects. In this case, depicted in figure 2, only 11 percent of the sample would not receive tenure. Here we see that, of those excluded, almost 60 percent are in the lowest two performance quintiles in reading post-tenure, and a full 45 percent are in the very lowest quintile. In math, of those excluded, nearly 70 percent are in one of the two lowest quintiles of post-tenure performance, with nearly half falling into the very lowest quintile. Whether the trade-offs we describe in figures 1 or 2 are worthwhile is no doubt a judgment call, which ultimately depends on what one considers the relevant alternatives to be.

CONCLUDING THOUGHTS: IN THE EYE OF THE BEHOLDER

What does all this mean for tenure policy? We suspect the results presented in this brief will tend to reinforce views on both sides of the policy divide over whether VAM estimates of teacher job performance ought to be used for high-stakes purposes like determining tenure. Those opposed to the idea might point to the finding that the year-to-year correlations in teacher effects are modest, and that we cannot know the extent to which this reflects true fluctuations in performance or changes in class or school dynamics outside of a teacher's control (such as the oft-mentioned dog barking outside the window on testing day). Moreover, in the case of reading effects, we predict that a non-trivial percentage of teachers who are found to be ineffective pre-tenure appear to be more effective in a post-tenure period.

On the flip side, supporters of VAM-based reforms might note that these inter-temporal estimates are very much in line with findings from other sectors of the economy that do use them for policy purposes. Perhaps more importantly, based simply on the percentage of teachers who move from the 4th to 5th year, it does not appear that schools are very selective in terms of which teachers receive tenure. Nevertheless, pre-tenure estimates of teacher job performance clearly do predict estimated post-tenure performance in both subjects, and would therefore seem to be a reasonable metric to use as a factor in making substantive teacher selection decisions.

It is important to note that policies relying on the use of value-added measures of teacher effectiveness depend on the accuracy, precision, and stability of the estimates, and, to date, there has been relatively little research on these issues. And, while research on the use of VAM to predict *individual* teacher performance is very much in its infancy, the extent to which model specification affects individual estimates of teacher job performance is a research topic that is finally receiving much-needed attention. Of particular importance is whether floor and ceiling effects or test content influence teacher-effect estimates (Koedel and Betts 2008; Lockwood et al. 2008); whether teachers are equally effective across

different student subgroups or teaching contexts (Goldhaber and Hansen 2008); whether VAM measures of teacher job performance can be validated based on experiments (Kane and Staiger 2008); and the extent to which VAM estimates line up with other ways of assessing teacher job performance (Jacob and Lefgren 2005).

Beyond these issues is the question of whether a change in pay or job retention policies governing the teacher labor market might significantly change teacher behavior. Using VAM to inform pay or tenure decisions would represent a seismic shift in teacher policy. Such a shift could have far reaching consequences for who opts to enter the teacher labor force and how teachers in the workforce behave. Teaching, for instance, is a fairly risk-free occupation in the sense that salary is currently governed by degree and experience levels, and such jobs are relatively secure. Policies that make the occupation more risky might induce different types of entrants, but economic theory would also suggest that teacher quality would only be maintained if salaries were increased enough to offset any increased risk associated with becoming a teacher. Finally, some evidence has shown that the stability of job performance may increase in the presence of incentive systems (Judeisch and Schmidt 2000; Roth 1978). All of this suggests that we likely cannot know the full impact of using VAM-based reforms without conducting assessments of actual policy variation.

REFERENCES

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 25(1): 95–135.

Ballou, D. 2005. "Value-Added Assessment: Controlling for Context with Misspecified Models." Paper presented at the Urban Institute Longitudinal Data Conference, March.

Ballou, D., W. Sanders, and P. Wright. 2004. "Controlling for Student Background in Value-Added Assessment of Teachers." *Journal of Educational and Behavioral Statistics* 29(1): 37–66.

Boyd, D., P. Grossman, H. Lankford, S. Loeb, and J. Wyckoff. 2005. "How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement." National Bureau of Economic Research Working Paper 11844.

Boyd, D., H. Lankford, S. Loeb, and J. Wyckoff. 2005. "Explaining the Short Careers of High-Achieving Teachers in Schools with Low-Performing Students." *American Economic Review* 95(2).

Clotfelter, C., H. Ladd, and J. Vigdor. 2006. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." *Journal of Human Resources* 41(4): 778–820.

Deadrick, D. L., and R.M. Madigan. 1990. "Dynamic Criteria Revised: A Longitudinal Study of Performance Stability and Predictive Validity." *Personnel Psychology* 43:717–44.

Goldhaber, D. 2006a. "Everyone's Doing It, but What Does Teacher Testing Tell Us about Teacher Effectiveness?" CALDER working paper.

———. 2006b. "National Board Teachers Are More Effective, but Are They in the Classrooms Where They're Needed the Most?" *Education Finance and Policy* Summer 1(3).

Goldhaber, D., and E. Anthony. 2007. "Can Teacher Quality Be Effectively Assessed? National Board Certification as a Signal of Effective Teaching." *Review of Economics and Statistics* 89(1): 134–50.

Goldhaber, D. and M. Hansen. 2008. "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance." CRPE Working Paper #2008-5. Available online at www.crpe.org.

Gordon, R., T. Kane, and D. Staiger. 2006. "Identifying Effective Teachers Using Performance on the Job." Hamilton Project White Paper 2006-01.

Hanushek, E.A. Forthcoming. "Teacher Deselection." In *Creating a New Teaching Profession*, edited by Dan Goldhaber and Jane Hannaway. Washington, DC: Urban Institute Press.

Hanushek, E., J. Kain, and S. Rivkin. 2004. "Why Public Schools Lose Teachers." *Journal of Human Resources* 39(2): 326–54.

Hanushek, E. A., J. Kain, D. O'Brien, and S. Rivkin. 2005. "The Market for Teacher Quality." National Bureau of Economic Research Working Paper 11124.

Hoffman, David A., Rick Jacobs, and Steve J. Gerras. 1992. "Mapping Individual Performance Over Time." *American Psychological Association* 77(2): 185–95.

Hoffman, David A., Rick Jacobs, and Joseph E. Baratta. 1993. "Dynamic Criteria and the Measurement of Change." *Journal of Applied Psychology* 78(2): 194–204.

Jacob, B., and L. Lefgren. 2005. "Principals as Agents: Subjective Performance Measurement in Education." National Bureau of Economic Research Working Paper 11463.

Joyce, Robert P. 2000. "The Law of Employment in North Carolina's Public Schools." Institute of Government, University of North Carolina Chapel Hill. Accessed 11/17/08 at http://www.sog.unc.edu/pubs/electronicversions/pdfs/leps18.pdf.

Judiesch, Michael K., and Frank L. Schmidt. 2000. "Between-Worker Variability in Output under Piece-Rate Versus Hourly Pay Systems." *Journal of Business and Psychology* 14(4): 529–52.

Kane, T., and D. Staiger. 2001. "Improving School Accountability Measures." National Bureau of Economic Research Working Paper 8156.

———. 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives* 16(4): 91–114.

———. 2008. "Are Teacher-Level Value-Added Estimates Biased? An Experimental Validation of Non-Experimental Estimates." Working Paper.

Kane, T., J. Rockoff, and D. O. Staiger. 2006. "What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City." Working Paper.

Kane, T., D. Staiger, and J. Geppert. 2002. "Randomly Accountable." *Education Next* 2(1): 57–61.

Koedel, C., and J. Betts. 2007. "Re-examining the Role of Teacher Quality In the Educational Production Function." Working Paper 0708, Department of Economics, University of Missouri.

———. 2008. "Value-Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation." National Center on Performance Incentives Working Paper 2008-21.

McCaffrey, D. F., T. Sass, and J.R. Lockwood. 2008. "The Intertemporal Stability of Teacher Effect Estimates." National Center on Performance Incentives Working Paper 2008-22.

McCaffrey, D., D. Koretz, J. Lockwood, and L. Hamilton. 2004. *Evaluating Value-Added Models for Teacher Accountability*. Santa Monica, CA: RAND Corporation.

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

Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Students' Achievement: Evidence from Panel Data." *American Economic Review* 94(2): 247–52.

Roth, H. F. 1978. "Output Rates among Industrial Employees." *Journal of Applied Psychology* 63:40–46.

Rothstein, J. 2008a. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." National Bureau for Economic Research Working Paper 14442.

———. 2008b. "Student Sorting and Bias in Value Added Estimation: Selection on Observables and Unobservables." June.

Sanders, W., J. Ashton, and S. Wright. 2005. "Comparison of the Effects of NBPTS Certified Teachers with Other Teachers on the Rate of Student Academic Progress." Final report requested by the National Board for Professional Teaching Standards.

Sass, T. R. 2008. "The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy." Policy Brief. Washington, DC: Urban Institute, Center for Analysis of Longitudinal Data in Education Research.

Todd, P. E., and K. I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal* 113:F3–F33.

Zevin, J. 1974. "In Thy Cooperating Teacher's Image: Convergence of Social Studies Student Teachers' Behavior Patterns with Cooperating Teachers' Behavior Patterns." Education Resources Information Center. ERIC#: ED087781.

NOTES

[1] See, for instance, Hanushek et al. (2004); Kane et al. (2006); and Rockoff (2004) for a discussion of general variation in effectiveness; Boyd, Grossman, et al. (2005); Goldhaber (2006a); and Kane et al. (2006) for a discussion of variation by licensure area and licensure test performance; and Goldhaber (2006b) and Sanders et al. (2005) by variation by NBPTS certification status.

[2] There is a growing body of literature that examines the implications of using value-added models (VAMs) in an attempt to identify causal impacts of schooling inputs and indeed, the contribution that individual teachers make toward student learning gains (Ballou 2005; Ballou et al. 2004; Kane and Staiger 2008; McCaffrey et al. 2004; Rivkin 2004; Rothstein 2008a; Todd and Wolpin 2003).

[3] See Aaronson et al. (2007); Ballou, (2005); Goldhaber and Hansen (2008); Koedel and Betts (2007); and McCaffrey et al. (2008).

[4] More extensive findings, along with greater detail about the data and methodology used to estimate the findings reported here, are available in Goldhaber and Hansen (2008)—a new CRPE Working Paper (#2008-5) that can be found at www.crpe.org.

[5] Recent research illustrates how these data can be used for analyzing the effects of schools and teachers on students (Clotfelter et al. 2006; Goldhaber and Anthony 2007; Goldhaber, 2006a, 2006b).

[6] If a teacher takes disability, sick, or medical leave in the pre-tenure period, and does not work 120 days in one of the years, that year does not count toward tenure, nor does it reset the tenure clock. This is a state policy—discretion is not left to the district. However, if tenure is obtained in another district in NC and a teacher changes districts, the new district has discretion to either grant tenure immediately, after Year 1, or vote to observe one more year and grant it after Year 2 in the district.

[7] A teacher may stay in the workforce, but would only remain in our sample if they stayed teaching at the 5th grade level.

[8] Rothstein (2008b) shows the VAM specification with zero to minimal bias is that which includes student background covariates and a full history of test scores; thus, we primarily use this model specification here.

Rothstein's conclusion stems from the significance of this vector in explaining variation in student achievement. This vector captures almost all of the variation in student achievement, leaving very little room for non-random sorting or the likelihood to bias student estimates.

[9] Specifically, we estimate:
$A_{i,j,t,s,g=5} = \alpha \mathbf{A}_{\mathbf{i(history)}} + X_{i,t,g=5}\gamma + \tau_{j,t,g=5} + \varepsilon_{i,j,t,s,g=5}$ where $\mathbf{A}_{\mathbf{i(history)}} = \left[A_{i,R,g=4} \middle| A_{i,M,g=4} \middle| A_{i,R,g=3} \middle| A_{i,M,g=3}\right]$
where i represents students, j represents teachers, k represents schools, s represents subject area (math or reading), and t represents the school year. Student achievement, A_{ijkst}, is regressed against: prior student achievement, $A_{ijks(t\text{-}1)}$; a vector of student and family background characteristics (for example, age, race and ethnicity, disability and free or reduced-price lunch status, parental education level), X_{it}; a vector of classroom characteristics (such as class size or average student characteristics, including achievement), C_{jt}; and teacher, τ_j, school, ς_k, and year, ϕ_t, fixed effects. The error term is associated with a particular student in a particular year ε_{ijkst}.

[10] This is moderately higher than the general level of variation reported among elementary school teachers in McCaffrey et al. (2008), though slightly lower than the levels found among secondary teachers.

[11] The post-tenure job performance estimates are based solely on teacher effectiveness estimates for teachers in experience years 5 and over. Thus, we have purposefully excluded experience years 3 and 4 (in the case of the two-year teacher effect estimates) and year 4 (in the case of the three-year teacher effect estimates). However, other regressions using the first 3 years versus the first 4

years of pre-tenure performance to predict post-tenure performance yield similar results.

[12] The correlation of the underlying pre- and post-tenure estimated teacher effects rises from 0.34 to 0.40 when moving from the two-year pre-tenure effects estimates to the three-year estimates.

[13] The correlation between pre- and post-tenure estimated teacher effects is 0.48 with the two-year estimates and 0.56 with the three-year estimates.

[14] It is difficult to tell how job performance assessments influence teacher retention pre-tenure, since an unknown number of teachers are likely counseled out of the profession in ways that may not show up on formal documentation.

[15] In North Carolina, teachers are typically awarded tenure after four successive years of teaching, though the specific requirements vary somewhat and depend on whether a teacher has been tenured in another state and/or has prior teaching experiences in other states. Thus, we are not precisely sure which teachers in our data actually receive tenure, but make the assumption that it is all teachers who continue teaching after four successive years of employment in North Carolina public schools. The requirements for achieving tenure in North Carolina are described in Section 1803 of the School Employment Law (Joyce 2000).

ABOUT THE AUTHORS

Dan Goldhaber is Research Professor at the University of Washington Bothell and the Center on Reinventing Public Education (CRPE). Dr. Goldhaber is the principal investigator of the CALDER Washington team. Dr. Goldhaber's research focuses on issues of educational productivity and reform at the K–12 level, and the relationship between teacher labor markets and teacher quality.

Michael Hansen is a PhD student in economics at the University of Washington and a research assistant at the Center on Reinventing Public Education. His research primarily deals with public school teacher quality, performance measurement, and behavioral changes in response to policy changes.

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

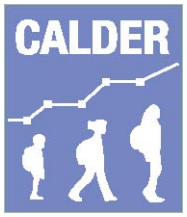
IN THIS ISSUE

Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions

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.

The research presented here is based primarily on confidential data from the North Carolina Education Research Center (NCERDC) at Duke University, directed by Clara Muschkin and supported by the Spencer Foundation. The authors wish to acknowledge the North Carolina Department of Public Instruction for its role in collecting this information. They gratefully acknowledge the Institute of Educational Studies at the Department of Education for providing financial support for this project. This paper has benefited from helpful comments from participants at the APPAM 2008 Fall Research Conference, the University of Virginia's 2008 Curry Education Research Lectureship Series, and the Economics Department Seminar Series at Western Washington University. The authors also wish to thank Carol Wallace for editorial assistance. Responsibility for any and all errors rests solely with the authors.