

THE INTER-TEMPORAL VARIABILITY OF TEACHER EFFECT ESTIMATES*

Daniel F. McCaffrey
The RAND Corporation
4570 Fifth Avenue, Suite 600
Pittsburgh, PA 15213
Voice: 412-683-2300 X4919
Email: danielm@rand.org

J. R. Lockwood
The RAND Corporation
4570 Fifth Avenue, Suite 600
Pittsburgh, PA 15213
Voice: 412-683-2300 X4941
Email: lockwood@rand.org

Tim R. Sass
Florida State University
Department of Economics
Tallahassee, FL 32306-2180
Voice: 850-644-7087
Email: tsass@fsu.edu

Kata Mihaly
The RAND Corporation
1200 South Hayes Street
Arlington, VA 22202-5050
Voice: 703-413-1100 X5393
Email: kmihaly@rand.org

*We thank the Editor and an anonymous reviewer for helpful comments on an earlier draft of the manuscript. Ryan Murphy and Micha Sanders provided able research assistance. This paper has not been formally reviewed and should not be cited, quoted, reproduced, or retransmitted without the authors' permission. This material is based on work supported by a supplemental grant to the National Center for Performance Initiatives funded by the United States Department of Education, Institute of Education Sciences. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of these organizations.

Abstract

School districts have begun using estimates of teachers' effects on student test scores (or "value-added") for diagnostic purposes and allotting monetary rewards. Such estimates 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.

I. Introduction

There is growing interest in using student outcomes to evaluate teachers when making decisions about teacher retention and compensation. For any performance-based personnel system to provide the correct incentives and enhance teacher quality, it is necessary that there be a strong link between true performance and reward or retention. Thus, at any point in time it is necessary that measures of teacher performance provide an accurate (unbiased) measure of teacher productivity. Avoiding systematic errors in evaluating teacher performance is not sufficient, however. If outcome-based measures of teacher quality are unbiased, yet highly variable, their efficacy in high-stakes personnel decisions will be limited. For example, there are proposals to use estimates of teachers' effects on student test scores or "value-added" to determine which teachers are granted tenure and which are dismissed after an initial probationary period.¹ 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 prove to be relatively ineffective in boosting achievement. Similarly, if variability in outcome-based 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 systems.

While the issues of bias and variability in estimates of teacher effectiveness are both important, and to some degree intertwined, in this paper we focus primarily on the within-teacher variability in estimated effectiveness over time and the associated implications for a viable outcome-based system of teacher personnel decisions.² First, we establish that estimates for

¹ See, for example, Gordon, Kane and Staiger (2006).

² For discussions of the ongoing controversy over the ability of statistical models to generate unbiased estimates of a teacher's effect on student achievement with non-experimental data see Andrabi, et al. (2008), Kane and Staiger (2008) and Rothstein (2008).

teachers with few students are very imprecise and we therefore restrict our subsequent analysis to teachers with at least 15 students in any given year. Using this subsample, we examine year-to-year correlations in estimated effects and relate these values to inter-temporal variation in the measured performance of workers in other occupations. We then decompose the variance in estimated teacher effects into three sources: persistent effects, non-persistent changes, and sampling error, and show that little of the non-persistent changes within teacher can be explained by observable time-varying teacher characteristics, such as experience, formal educational attainment and in-service training. We examine how the variance decomposition of the estimates depends on factors including the value-added model specification, the grade level (elementary versus middle), district, and the achievement test. Finally, using our decomposition we explore the effect of averaging teacher effect estimates over multiple years and consider the implications of this averaging for practical systems of teacher evaluation.

II. Prior Studies

Only a few previous studies have measured the variability of teacher effects and none have analyzed the sources of inter-temporal variation in depth. Ballou (2005) compares the rankings of elementary and middle school teachers in a “moderately large” Tennessee school district across two years. He finds that 40 percent of mathematics teachers who are ranked in the bottom quartile of teacher quality rankings in the first year remain in that quartile the following year and 30 percent move into the top two quartiles. At the other end of the quality distribution, nearly 50 percent of mathematics teachers in the top quartile in one year are also in the highest quartile the next year while roughly 30 percent fall into the bottom two quartiles. Ballou also shows that the precision of teacher effect estimates increases with the number of annual

observations per teacher. Estimating teacher effects over a three-year span, 58 percent of middle school math teachers have estimates significantly different from the average teacher effect whereas with single-year estimates only 30 percent of the estimated teacher effects for middle school math teachers are significantly different from the average.

Similarly, Aaronson et al. (2007) compare the rankings of estimated teacher effects for Chicago public school teachers across two years. They find that 36 percent of teachers ranked in the lowest quartile in the first year also rank in that quartile in second year, 29 percent move up to the second quartile and the remaining 35 move into the top half of the distribution. At the other end of the scale, 57 percent of the teachers in the top quartile in the first year remain there in year two. Another 23 percent move down to the third quartile and only 20 percent fall down into the lower half of the quality distribution.

Koedel and Betts (2007) conduct a similar analysis, comparing the ranking of San Diego teachers in two years based on their fixed-effect estimates. While a large fraction of teachers stays in the same quintile from one year to the next, the degree of persistence is less than that found by Aaronson, et al. (2007) in Chicago. Among teachers who are ranked in the lowest quintile in the first year, 30 percent stay in that quintile, but a nearly equal proportion (31 percent) move into the top two quintiles in the second year. Similarly, 35 percent of teachers initially ranked in the top quintile remain there in the second year while 30 percent fall into the first or second quintiles of the quality distribution in year two. These comparisons are based on estimates of within-school teacher effects (i.e., achievement models that include student, teacher and school fixed effects). Omitting student and school fixed effects, they find the teacher effects to be more stable; 43 percent of teachers in the bottom quintile stay there in the next year and 50 percent of teachers in the top quintile in the first year remain there in the second year.

Our study builds on existing work on the variability of teacher value-added measures by decomposing the variance in teacher effect estimates into persistent components, non-persistent changes and sampling errors, analogous to the decomposition used by Kane and Staiger (2002) for school-level performance measures. This decomposition permits examination of how much averaging measures over time would reduce inter-temporal variability, and also provides a framework for characterizing how different value-added model specifications may contain different amounts of systematic and non-systematic errors. Also, we conduct our investigations using five large county-level districts in Florida, allowing us to examine the sources of variance in teacher performance measures across districts, grade levels (elementary and middle school), different achievement tests, and different value-added models.³

III. Methods

A. Estimation of Teacher Effects

In order to construct an empirical model of teacher effects, we begin with the cumulative model of student achievement developed by Boardman and Murnane (1979) and Todd and Wolpin (2003). Following these authors, we assume: (i) a cumulative achievement function that is invariant to a student's age, linear and additively separable and (ii) children are endowed at

³ Subsequent to the initial version of this paper, Goldhaber and Hansen (2008) produced a complementary analysis of the stability of measured teacher effects in North Carolina. While their study differs from ours in a number of ways, including their focus on elementary school teachers and the way in which they decompose the variance in teacher effects, their findings tend to reinforce the conclusions we draw from our analysis of estimated teacher effects in Florida. In particular, they find year-to-year correlations in teacher value-added are moderate (0.3 in reading, 0.5 in math), with a large proportion of the year-to-year variation due to sampling error. The sampling error diminishes with increases in the number of students per teacher, yet even for teachers who have many students there is significant within-teacher variation over time that is not explained by observable teacher characteristics.

birth with a fixed level of ability and parental inputs that do not change over time. Given these assumptions we can denote a student's achievement level at age t as:

$$A_{it} = \sum_{h=1}^t [\beta_{0h} + \beta_{1h} \mathbf{X}_{ih} + \beta_{2h} \mathbf{P}_{-ijmh} + \beta_{3h} \mathbf{T}_{kh} + \beta_{4h} \mathbf{S}_{mh}] + \beta_{5h} \mu_{0i} + \eta_{it} \quad (1)$$

The current achievement level is a function of all current and prior school-based inputs, \mathbf{X} , \mathbf{P} , \mathbf{T} , \mathbf{S} , the initial ability/parental-input endowment, μ_{0i} , and measurement error in the test of achievement, η_{it} . The vector \mathbf{X}_{ih} represents student-specific school-based inputs, such as participation in a specific program. Classroom peer characteristics are represented by the vector \mathbf{P}_{-ijmh} where the subscript $-i$ denotes students other than individual i in classroom j in school m . Teacher inputs are represented by a vector of characteristics for the teacher (where k indexes teachers) who teaches the child at each age, \mathbf{T}_{kh} . The school-level input vector is denoted \mathbf{S}_{mh} .

If the marginal impacts of all school-based inputs (\mathbf{X} , \mathbf{P} , \mathbf{T} , \mathbf{S}) decline geometrically over time, current achievement can be represented by:⁴

$$A_{it} = \lambda A_{it-1} + \beta_{0t} + \beta_{1t} \mathbf{X}_{it} + \beta_{2t} \mathbf{P}_{-ijmt} + \beta_{3t} \mathbf{T}_{kt} + \beta_{4t} \mathbf{S}_{mt} + (\beta_{5t} - \lambda \beta_{5t-1}) \mu_{0i} + \eta_{it} - \lambda \eta_{it-1} \quad (2)$$

Current achievement depends only on contemporaneous school-based inputs, the initial endowment, and an error term. The lagged test score, A_{it-1} , serves as a sufficient statistic for all prior school-based inputs.⁵

⁴ See Boardman and Murnane (1979), Todd and Wolpin (2003) or Harris and Sass (2006) for a detailed derivation.

⁵ If the impact of student/family inputs diminishes at the same rate as prior schooling inputs then the lagged test score can serve as a sufficient statistic for both lagged schooling and student/family inputs. In this case, the initial endowment term, $\omega \mu_{0i}$, drops out of the equation.

Assuming the marginal effect of the initial endowment declines at a constant rate then $(\beta_{5t} - \lambda\beta_{5t-1})$ is a constant and can be denoted by ω . We can also let $\varphi_{it} = \eta_{it} - \lambda\eta_{it-1}$ and $\beta_{3t}\mathbf{T}_{kt} = \delta_{kt}$. This yields a general “value-added” model:

$$A_{it} = \lambda A_{it-1} + \beta_{0t} + \beta_{1t}\mathbf{X}_{it} + \beta_{2t}\mathbf{P}_{-ijmt} + \beta_{4t}\mathbf{S}_{mt} + \omega\mu_{0i} + \delta_{kt} + \varphi_{it} \quad (3)$$

Commonly estimated forms of equation (3) vary in their assumptions regarding the persistence of prior school-based inputs, λ , and modeling of the fixed student/family endowment effect on current achievement, $\omega\mu_{0i}$. The persistence of prior-year inputs can either be estimated from the model or one can assume complete persistence of prior inputs (i.e., $\lambda=1$).⁶ In the latter case the lagged test score can be subtracted from both sides, yielding the change in student achievement, $\Delta A_{it} = A_{it} - A_{it-1}$, as the dependent variable. The effect of student endowments is either captured by observable time-invariant student characteristics, \mathbf{Z}_i , (so that $\omega\mu_{0i}$ is assumed to be fully described by $\beta_5\mathbf{Z}_i$, for an unknown vector of coefficients β_5) or by a student fixed effect, $\gamma_i = \omega\mu_{0i}$ (which accounts for both observed and unobserved student characteristics). Following Rothstein (2008), we consider three variants of equation (3), based on differing assumptions about persistence and alternative treatments of student heterogeneity:

$$\Delta A_{it} = \beta_{0t} + \beta_{1t}\mathbf{X}_{it} + \beta_{2t}\mathbf{P}_{-ijmt} + \beta_{4t}\mathbf{S}_{mt} + \mathbf{Z}_i + \delta_{kt} + \varphi_{it} \quad (4A)$$

$$A_{it} = \lambda A_{it-1} + \beta_{0t} + \beta_{1t}\mathbf{X}_{it} + \beta_{2t}\mathbf{P}_{-ijmt} + \beta_{4t}\mathbf{S}_{mt} + \beta_5\mathbf{Z}_i + \delta_{kt} + \varphi_{it} \quad (4B)$$

⁶ Including the lagged test score as an explanatory variable is potentially problematic. Since v_{it} is a function of the lagged error, η_{t-1} , the lagged achievement term, A_{it-1} , will be correlated with the error term in equation (3), and OLS estimates of equation (3) will in general be biased except in the unlikely case that the error terms are also autocorrelated with correlation λ . This potential bias does not exist if persistence is complete because the lagged test score is on the left hand side of the equation and is no longer an explanatory variable.

$$\Delta A_{it} = \beta_{0t} + \beta_{1t} \mathbf{X}_{it} + \beta_{2t} \mathbf{P}_{-ijmt} + \beta_{4t} \mathbf{S}_{mt} + \gamma_i + \delta_{kt} + \varphi_{it} \quad (4C)$$

In each value-added specification the teacher-by-year effect, δ_{kt} , represents the average achievement of a teacher's students in a given year, conditional on prior school inputs, student/family endowments and contemporaneous non-teacher schooling inputs (e.g., classroom peers and school-level factors such as school leadership). Note that since the models do not include school fixed effects, the teacher effect is measured relative to the average of all teachers in the relevant subject, grade range and jurisdiction, not the average teacher at a given school.⁷

B. Sources of Inter-temporal Variability in Estimated Teacher Effects

Following the analysis of school-level average achievement by Kane and Staiger (2002), we can identify two sources of variation over time in the annual teacher effect estimates: sampling error and non-persistent changes in performance. Sampling error refers to errors in the estimated teacher effects due to idiosyncrasies in individual student scores, after controlling for other factors in the model. The idiosyncratic outcomes for individual students tend to average out across a teacher's students and thus the sampling error will tend to fall within the number of students per teacher per year. The standard errors on the estimated teacher effects measure the contribution of sampling error to inter-temporal instability in estimated effects.

⁷ The default solution for estimating teacher fixed effects in most statistical software packages is to contrast all teachers to an arbitrary holdout teacher, such as the teacher with the highest or lowest identification number in the data set. As described in Mihaly, et al. (2009), the selection of the holdout teacher does not affect the relative rankings of teachers within a grade level and year, but it does alter the absolute teacher effect estimates. Contrasting teachers to different arbitrary holdout teachers each year could lead to considerable inter-temporal variability in any given teacher's estimated effects. *Post-hoc* centering of effects could remove this source of inter-temporal variation but then the estimated standard error of the teacher fixed effect estimates would be wrong and computationally challenging to correct with large samples of teachers. As an alternative, we contrast all teachers to the average teacher effect within a given year and grade level. To accomplish this we employ the Stata program `felsdm`. See Mihaly, et al. (2009) for details on the development and use of `felsdm`.

Non-persistent changes in performance refer to all sources of year-to-year changes in the estimated impact of a teacher on student achievement, other than sampling errors. These might include variation in the teacher's true performance, "chemistry" between students within a class, the impact of a disruptive student, test day conditions, matches between the specific test items and the concepts emphasized by the teacher or any other classroom-level factors that vary across years. As noted in Kane and Staiger (2002), unlike sampling error, we cannot directly estimate the variability of non-persistent changes in performance. Rather, we do so indirectly by comparing the average variability among estimated effects within-teacher with the variability due to sampling error. If the variability in estimated effects exceeds the variability due to sampling error then the remaining variance is attributed to non-persistent changes.

The persistent teacher effect is simply the portion of the estimated effect that is common across years. It is not necessarily equal to the teacher's true performance; estimated effects from value-added models might not equal true causal effects of teachers due to violations of the model assumptions (c.f., Rothstein (2008)) and even persistent components of the estimated effects might include confounding factors that are persistent across years rather than causal effects. For example, if the achievement model fails to properly capture all unobservables that are correlated with classroom assignments and the classroom average of the unobservables is stable across years, then these omitted variables will be part of the persistent teacher effect. In an extreme case of confounding, suppose annual teacher effects were measured by classroom average test scores without any adjustment. These effects would likely demonstrate strong persistence within-teacher over time due to the stability in the types of students assigned to teachers across years.

As discussed by Kane and Staiger (2002) in the context of school-level effects, the usefulness of teacher effect estimates depends on the variance in sampling errors and the

variance in non-persistent changes relative to the variability across teachers in the persistent effects. Large variance in non-persistent change and sampling error, relative to persistent teacher effects, leads to low correlation of estimated effects between adjacent years. The year-to-year correlation in teacher effects is roughly equal to the ratio of the variance in persistent effects to total variance (i.e., the sum of variances in sampling error, non-persistent effects and persistent effects). Teacher effect estimates that exhibit low year-to-year correlations have limited utility because they fail to yield information that is stable enough to support decisions about teachers.

By separating the sampling errors from the non-persistent change we determine if the only source of inter-temporal instability is sampling error or if other sources also contribute, possibly including true variation in teacher performance. Large variation in true performance across years would suggest that pooling data across years to smooth out could result in bias and obscure important changes in performance. Also, decomposing the variance of teacher effects provides insights into the relative utility of alternative achievement model specifications. Two different value-added models might yield teacher effect estimates with similar year-to-year correlations, but may have very different variance components, which in turn can provide indirect evidence on the extent of bias in the alternative models.

C. Variance Decomposition and Stability Metrics

In this section we formalize the three components of the variance in the estimated teacher effect with a simple statistical model and describe how we use the model to estimate the variance components. We then use the estimated variance components to define a measure of inter-temporal variation in estimated teacher effects, which we call the “stability coefficient.”

We consider the following model for the estimated effects δ_{kt} for teacher k in year t :

$$\delta_{kt} = \theta_k + \xi_{kt} + \varepsilon_{kt}. \quad (5)$$

We assume the teacher-by-year effect, δ_{kt} , can be represented by the sum of three independent random variables: θ_k , the teacher’s persistent effect (that has mean zero and unknown variance τ^2); ξ_{kt} , the non-persistent changes in the teacher’s performance (that has mean zero and unknown variance v^2); and ε_{kt} , the sampling errors (with mean zero and variance se_{kt}^2 , equal to the square of the standard errors in the estimated teacher effects). To estimate the unknown variance components we make the additional assumption that the random variables are normally distributed and maximize the resulting likelihood for the observed estimates pooling the data within and among teachers.⁸

The estimated variance components allow us to develop several metrics to help characterize the various estimators of teacher effects as measures of teacher performance. As discussed above, for some purposes it is useful to distinguish sampling error from the other sources of variance in the estimated effects (persistent effects and non-persistent changes), which combined we refer to as the “annual signal.” The ability to differentiate teacher performance in a given year depends on the ratio of variability in the annual signal to total variability in the estimates and is called the *reliability coefficient*,

$$Reliability = \frac{\tau^2 + v^2}{\tau^2 + v^2 + se_{kt}^2}.$$

The stability of estimates across years depends on the proportion of variability in the estimates that is due to the persistent effects, so we define the *stability coefficient* as

⁸ To improve estimation we estimate the natural log of τ^2 and the natural log v^2 and transform the resulting estimates to the original scale.

$$Stability = \frac{\tau^2}{\tau^2 + \nu^2 + se_{kt}^2}.$$

The ratio of the stability coefficient of the measure over time to its reliability as a measure of annual performance is $\tau^2/(\tau^2 + \nu^2)$, the proportion of annual signal variance that is between teachers, or equivalently, the proportion of annual signal variance that is due to the persistent effects. Analogously $\nu^2/(\tau^2 + \nu^2)$ is the proportion of the annual signal variance that is due to non-persistent changes. Both *Reliability* and *Stability* depend on the standard errors of individual teacher estimates and we summarize these with averages across teachers. We can estimate both *Stability* and *Reliability* using the estimates of τ^2 and ν^2 . Alternatively, as shown in the Technical Appendix, under our model for the estimated effects, the cross-year correlation for estimated effects from adjacent years is also equal to *Stability*. Because the standard errors differ across estimates, the two methods of estimating *Stability* will differ somewhat, but both can be interpreted as measures of stability.

As discussed in Lockwood, Louis, and McCaffrey (2002), *Reliability* is the primary determinant of the ability to distinguish among, and accurately rank, teachers in a single year. *Stability* plays a similar role for distinguishing among teachers' future performance and for creating rankings that are stable across time. If the data are normally distributed and our model holds, then the stability coefficient determines the proportion of teachers who will switch quintile ranks across years and how much error will be in predictions of future teacher effects. For example, if the estimated effects have a stability of 0.3 then we can expect that about 24, 19, 15, and 10 percent of teachers ranked in the first quintile in one year will be ranked in the second, third, fourth and fifth quintiles in the next year. For an estimated effect with a stability of 0.7 the

percentages are 26, 12, 5, and 1. The Technical Appendix provides details on the relationship between *Stability* and quintile rankings.

The stability coefficient also equals the relative reduction in prediction error variance due to using an estimated teacher effect (or average of multiple years of effects) to predict a teacher's future performance (i.e., the persistent effect). To see this, note that without any additional information, the uncertainty in predicting a teacher's persistent effect is the total variance in the persistent effect (τ^2). However, if we use the current single-year estimate to predict the persistent effect, then the variance of the prediction error is given by the variance in the persistent effect conditional on knowing the current value:

$$\text{Var}(\text{prediction error}) = \text{Var}(\text{persistent effect} \mid \text{current single-year estimate}) = \tau^2(1 - \tau^2/v^2).$$

The reduction in prediction error is thus equal to $\tau^2 - \tau^2(1 - \tau^2/v^2) = \tau^4/v^2$ and relative to the total error, the relative reduction in prediction error is $\tau^2/v^2 = \textit{Stability}$. The relative reduction in prediction error is analogous to the R^2 from linear regression, which provides another useful tool for calibrating the stability of the value-added measures.

Our measure of stability is also directly related to the efficiency of policies that use estimated teacher effects. As discussed in the Technical Appendix, a policy of tenuring only teachers whose estimated effect is above the 100 p th percentile of the distribution of estimated effects would improve the average teacher's persistent effect by $\sqrt{\tau^2/v^2} \times \tau\omega(p)$, where $\omega(p)$ is a factor that depends only on the normal distribution and p . Moreover, as shown in the Appendix, the maximum gain we could obtain from observing the actual persistent teacher effect, rather than a noisy estimate, would equal $\tau\omega(p)$, so the square root of the stability directly measures the inefficiency due to employing noisy measures of teacher effectiveness.

An advantage of our stability metric is its natural generalization to the average of two years in a way that is directly comparable to the measure of stability of a single-year estimate. Given our decomposition, the variances of sampling errors and non-persistent change in a two-year average of estimates equal $se_{kt}^2/2$ and $v^2/2$, respectively (assuming the standard error is constant across years). Hence the stability of a two-year average equals

$$Stability_2 = \frac{\tau^2}{\tau^2 + (v^2 + se_{kt}^2)/2}$$

Again, we can use our estimates of the various variances to calculate this quantity. More generally we can extend this formula to averaging any number of years of estimates.

Although the decomposition of the variability in estimated teacher effects is intuitive, our model relies on some important assumptions. In particular, we assume that non-persistent change is uncorrelated over time within a teacher. This precludes drifts in teacher performance over years in which a teacher might have a general level of performance but in which performance over a few years might systematically vary from the general level. For instance, a teacher's performance might deviate from his or her general level of performance as he or she adjusted to a new curriculum or a new principal. We explored these models with our data and for some estimators and some counties there was evidence of drift, but generally models without drift fit the data as well or better than models with drift and thus we focus on models without drift for this paper. A study of drift in effects would be useful and models that allow for drift may be important in other contexts.

IV. Data and Sample Selection

To estimate the achievement models and associated teacher effects we utilize data from the Florida Education Data Warehouse (FL-EDW), an integrated longitudinal database that covers all public school students and teachers in the state of Florida.⁹ From this statewide database we select data from five large school districts in the state, Dade, Duval, Hillsborough, Orange and Palm Beach. Each of the five districts enrolled 100,000 or more students in the 2004/05 school year and was among the 20 largest school districts in the United States. In addition to lowering computational costs compared to working with data from the entire state, selection of these five large districts allows us to determine how the stability of teacher effects varies across school districts and facilitates comparisons with the previous single-district studies in California, Illinois and Tennessee mentioned above.

The Florida data link both students and teachers to specific classrooms at all grade levels. However, achievement tests are only administered in grades 3-10 and thus current and lagged achievement are only observed in grades 4-10. The linkage between course content and what is tested on statewide exams may not be as strong for all high school students as it is in elementary and middle school. We therefore focus our analysis on students in grades 3-8 and estimate teacher effects for elementary and middle school math teachers.¹⁰ We select math teachers for our analysis because most studies of student achievement find a stronger correlation between school inputs and student achievement in math than in reading.

⁹ Detailed descriptions of the Florida data are provided in Sass (2006) and Harris and Sass (2008).

¹⁰ Middle school math courses are defined as math courses in which 90 percent or more of the enrolled students take either the 6th, 7th or 8th grade math achievement exam.

The State of Florida administers two achievement tests. The “Sunshine State Standards” Florida Comprehensive Achievement Test (FCAT-SSS) is a criterion-based exam designed to test for the skills that students are expected to master at each grade level. It is a “high-stakes” test that is used to assign grades to schools for state accountability purposes and to measure individual student performance for retention decisions and high school graduation. In our application the scores are normed to have mean zero and standard deviation one for each grade and year.¹¹ The FCAT-SSS has been used in selected grades since the 1998/99 school year, but was not implemented in all grades 3-10 until the 2000/01 school year. The second test is the FCAT Norm-Referenced Test (FCAT-NRT), a version of the Stanford Achievement Test used throughout the country. Version 9 of the Stanford test (the Stanford-9) was used in Florida through the 2003/2004 school year. Version 10 of the Stanford test (the Stanford-10) has been used since the 2004/05 school year. To equate the two versions of the exams we convert Stanford-10 scores into Stanford-9 equivalent scores based on the conversion tables in Harcourt (2002). Although scores on the Stanford-9 are scaled to a single developmental scale, we norm them by grade and year in order to make them comparable to the normed FCAT-SSS scores and so teacher effects and estimated variance components can be interpreted relative to variability in student achievement. We rely primarily on the FCAT-NRT exam since it provides an additional year of data. However, we also make comparisons across the two exams to determine how test differences may affect measured teacher performance.

¹¹ The FCAT-SSS was not designed to be a single developmental scale across grades. While a developmental scale conversion for the FCAT-SSS has been developed based on one year in which overlapping questions were administered on different grade-level exams, we choose to standardize the FCAT-SSS scores to have a mean of zero and standard deviation one for each grade and year so that differences in scores from adjacent years measure changes in relative position in the distribution rather than raw changes in scale scores.

The available data cover school years 1995/1996 through 2004/2005. However, given that testing of math achievement in consecutive grades did not begin until the 1999/2000 school year (for the FCAT-NRT) and the need to account for both current and lagged test scores, our analysis is limited to the five-year period, 2000/01 through 2004/05.

To avoid problems of attribution when students receive math instruction from multiple teachers, we restrict our analysis of student achievement to elementary students in “self-contained” classrooms and middle school students taking only one math course. We also exclude students repeating a grade. However, all students enrolled in a course are included in the measurement of peer-group characteristics. To avoid atypical classroom settings and jointly taught classes we consider only courses in which 50 or fewer students are enrolled and there is only one “primary instructor” of record for the class. Finally, we eliminate charter schools from the analysis since they may have differing curricular emphases and student-peer and student-teacher interactions may differ in fundamental ways from traditional public schools.

Our data contain a relatively rich set of student, peer, teacher and school characteristics. Time-varying student variables include student mobility, measured by the number of schools a student attends within a year, whether a student engages in a “structural move” between years (one in which at least 30 percent of his fellow students in the same grade at the initial school move to the same school) and whether a student undergoes a “non-structural” move (where fewer than 30 percent of students in the same initial school and grade made the same move). When student covariates are used instead of student fixed effects to measure student heterogeneity we employ the following time-invariant (or nearly time-invariant) student variables: gender, race/ethnicity, free/reduced-price lunch status, gifted program participation, limited English proficiency program participation and indicators for students with

speech/language, learning, cognitive, physical, emotional and “other” disabilities. Five variables capture important elements of classroom composition: the proportion of classmates who are female, the proportion who are black, the proportion who changed schools from the previous year, the average age of classroom peers and the total number of students in the class. For teachers we observe their experience (captured by a set of six indicators representing 1-2, 3-4, 5-9, 10-14 15-24 and 25+ years of experience), their recent in-service professional development (non-content and content oriented training hours in each of the previous three years), educational attainment (captured by an indicator for possession of an advanced degree), and an indicator of whether or not they are fully certified or hold a temporary license. We do not include teacher variables in our models for estimating teacher effects since they might be a component of the teacher effect of interest. However, we use them in *post hoc* analyses to determine how much year-to-year fluctuations in these variables contribute to the non-persistent change in estimated effects. At the school level we have time-varying data on the experience of the principal in administrative positions, the principal’s experience squared and whether the principal is in her first year as a principal at the school.

To ensure consistency in the samples used to estimate the three achievement model specifications, we restrict the sample to only those students with non-missing data on all of the student, peer, teacher and school variables and who possess at least two achievement gain scores (required for the model with student fixed effects). For analyses comparing the FCAT-SSS and FCAT-NRT exams we exclude any observations that lack valid data on both exam scores, ensuring comparability in the estimation samples.

V. Results

A. Sampling Error and the Variance of Estimated Teacher Effects

As noted above, sampling errors for individual students tend to average out across a teacher's students and thus to the extent that sampling error contributes to variability in teacher effect estimates, variance should fall within the number of students per teacher per year. The effect of the number of tested students per teacher on the precision of the teacher effect estimates is illustrated in Figure 1 and Table 1. Table 1 gives the average standard error by numbers of students used in estimating the teacher effects by county, grade level (elementary or middle) and model. For each county and grade level there are three rows, corresponding to teacher effect estimates derived by estimating equations 4A-4C. Figure 1 summarizes Table 1 by pooling data across counties.

The mean standard errors of the teacher effects for teachers with fewer than five students are very high, ranging from .46 to .92 depending on models and grade levels. Since the test scores are normed to have mean zero and standard deviation of one, a standard error of the teacher estimate of .5 indicates that even if teachers had no true effects and all the variability among teachers were sampling error, the variability among teachers would equal about 25 percent of the variance among students. Hence estimates for teachers based on very few students will tend to be extremely unstable across time. However, the standard errors of the teacher effects uniformly decrease with the number of students per teacher. Given that the sample sizes diminish considerably above 15 students per teacher in elementary school, we utilize a 15-student-per-teacher threshold in the remainder of the analysis. While middle school teachers with 20 or more students exhibit much lower mean standard errors in their estimated teacher

effects, the chosen 15-student minimum makes little difference; since they teach multiple classes in a school day, most middle school teachers teach more than 20 students in a school year.

Generally, the mean standard errors are similar between middle and elementary school teachers, except for the teachers with 20 or more students because middle school teachers in this group tend to have more students than elementary school teachers. Among models that use student covariates to control for student ability and family inputs, the standard errors of estimated teacher effects are insensitive to whether or not we assume complete or partial persistence of prior inputs. This suggests that either subtracting the prior score from the current score or controlling for it via regression accounts for about equal amounts of the variability in student scores.

The estimated teacher effects from the model with student fixed effects and complete persistence of past schooling inputs exhibit much higher average standard errors than the estimated effects from the other models. This is not surprising, given that there tends to be considerable variation among gain scores within students which suggest that neither student fixed effects nor time-invariant student covariates will explain a sizeable portion of the total variance in gain scores. Fixed effects, however, use many more degrees of freedom for explaining this variation in gain scores and these appear to be collinear with the teacher effects. This reduces the independent information we have for estimating teacher effects and consequently yields larger standard errors than do the models that employ student covariates to capture student heterogeneity. As described above, sampling errors are only one of two sources of inter-temporal variability among estimated effects – non-persistent change is the other – and how these two sources will play out among our different models and counties is explored through our study of the variability of estimated teacher effects.

We describe the variation over time in estimated teacher effects in two ways. First, in Table 2 we present correlations between estimated effects from adjacent years, broken down by grade level, for each of the five county-level school districts. In Table 3, we decompose the variance in estimated teacher effects from each model into the variance due to sampling error, non-persistent change, and persistent effects, again by model, county and grade-level. In Table 4, we demonstrate the implications of the inter-temporal variability of effects by exploring changes in quintile rankings of teachers based on estimated effects.

As shown in Table 2, estimates of cross-year correlations cover a wide range, but generally fall between 0.2 to 0.5 for elementary school teachers and 0.3 to 0.6 for middle school teachers. These values are consistent with those from prior studies of the inter-temporal variability of teacher effects in other jurisdictions. The difference between elementary school and middle school teachers is due in part to larger average number of students per teacher in middle schools and the resulting smaller standard errors in the estimated teacher effects. There are some differences across counties, but generally the estimates are similar and differences that do exist are not systematic.

There are, however, notable differences across models that are consistent across counties and grade-levels. In particular, the model with student covariates and partial persistence in past schooling inputs (Model 4B) yields estimates with the highest correlations in all cases. The cross-year correlations for estimates from this model tend to be about .10 to .14 higher than the correlations for other models in elementary school and about .23 higher than the correlations for other models in middle school. The difference in correlations of estimates between Model 4B and Model 4C (the model with student fixed effects and complete persistence in past schooling inputs) might be explained by the larger sampling errors in the estimates from Model 4C; as

shown in Table 1, the standard errors are on average about 75 to 80 percent larger for the estimates from Model 4C than for Model 4B. However, differences in the magnitude of sampling errors cannot be the source of the differences in the year-to-year correlation of estimates from Models 4A and 4B (the models with student covariates effects and differing persistence in past schooling inputs) because their average sampling errors are nearly identical.

For each model and county, Table 3 provides the average reliability (ratio of annual signal variance to total variance) across teachers and the proportion of the annual signal variance that is due to non-persistent change. Estimates from Model 4B have the largest reliability (i.e., sampling error accounts for the smallest share of the total variance) and Model 4C has the lowest reliability, which might be expected given the large sampling errors in the estimates from Model 4C (shown in Table 1). Also, relative to the estimates from Model 4A, non-persistent change accounts for a smaller proportion of the signal variance in estimates from Model 4B. The large reliability and relatively large proportion of signal variance due to persistent effects for estimates from Model 4B relative to Model 4A result because the variance of persistent effects is larger for the estimates from Model 4B and this leads to larger year-to-year correlations for Model 4B, even though the standard errors are roughly equal for the two models.

Another interesting feature of the decomposition presented in Table 3 is the relatively large size of the variance due to non-persistent change, especially for elementary school teachers. On average across counties, non-persistent change accounts for 30 to 40 percent of the signal variance of estimates for middle school teachers, depending on the model, whereas it accounts for between 46 and 54 percent of the signal variance of estimates for elementary teachers. These proportions are large and suggest that there is considerable year-to-year variability in teacher performance measures even after accounting for sampling error. However, this variance is not

explained by our observed time-varying teacher characteristics. County-level estimates of the percent of variance explained by these time-varying teacher characteristics were unstable, but analyses of pooled data from all five counties found that the percent of the variance in non-persistent effects explained by time-varying variables was less than one percent for elementary school teachers and two to seven percent, depending on the model, for middle school teachers.

The relatively weak stability of estimated effects across years would clearly have implications for using these effects for high-stakes personnel decisions. For example, rankings of teachers will tend to be unstable and rewards or sanctions based on ranks may be ineffective at achieving their goals. This is demonstrated in Table 4, which provides a tabulation of consecutive-year teacher rankings by quintile for each of the five school districts in the sample pooled across all years. Because of the instability of the estimated effects, only about one-third of teachers ranked in the top 20 percent one year are also ranked in the top quintile the following year and just half of the top-quintile teachers in a given year stay within the top two quintiles the next year. About 10 percent of these teachers are actually ranked in the bottom quintile the following year. Similarly, about one-third of the teachers in the lowest quintile in one year remain in the lowest quintile the next year, over half stay in the bottom two quintiles, and roughly 10 percent ranked in the top quintile the next year.

B. Stability of Performance in Other Professions

It is common to think of teacher performance as being relatively stable, i.e., a “good teacher” is always a “good teacher.” Thus at first blush the inter-temporal correlations and associated consistency in teacher rankings may seem “too low,” suggesting that the indirect method of evaluating teacher productivity through student test scores is leading to excessive variability in measured teacher performance. However, the inter-temporal correlations we obtain

are not out of line with those from other occupations where productivity can be measured more directly. Table 5 provides a summary of analyses of worker performance in other occupations including manufacturing employees, sales persons, university faculty, and professional athletes. As shown in the table, the correlation of most performance measures is modest at best, even for professions where first impression might suggest performance is very stable (major league hitters) and where very high-stakes decisions are based on the performance measures. Performance measures can be very highly correlated for individualized, repetitive tasks over short periods of time. For instance, the weekly output of piece-rate textile workers in contiguous weeks exhibits a correlation of about 0.9. However, even in these circumstances, when performance is compared over longer time spans, correlation in performance drops to about 0.55. For salespersons, university faculty, and baseball players, correlations in within-worker performance across months or years generally falls in the same range as we estimate for elementary and middle school teachers, about 0.2 to 0.7.

C. Differences in Stability Across Test Instruments

The analysis thus far has focused on estimated teacher effects based on student test scores from the Normed-Referenced Florida Comprehensive Assessment Test or FCAT-NRT. This is a “low-stakes” test in Florida, since it is not used in promotion decisions, teacher merit pay allocations or school grade assignments. One might expect that a different exam, such as Florida’s high-stakes criterion reference exam, the FCAT-SSS, would yield different results for three reasons. First, if different tests emphasize different kinds of material and the skills tested change more often for one test than another, then inter-temporal stability in estimated teacher effects can vary across tests. Second, different tests may have different effective maximums or “test ceilings.” A test with a low ceiling would tend to truncate scores for high-achieving

students and this might influence fluctuations in measured teacher performance over time. Third, there may be greater variation in teacher behavior over time with respect to a high-stakes test due to differential accountability pressure. If incentives to “teach to the test” vary over time with the degree of accountability pressure, this could increase the degree of measured variability in teacher performance over time.

As indicated in Table 6, using gains in the normed FCAT-SSS, rather than gains in the normed FCAT-NRT, does lead to differences in year-to-year correlations of teacher effectiveness, but there are no clear patterns that emerge. In some counties, the correlations are higher on FCAT-SSS than on the FCAT-NRT, whereas the opposite holds in other counties, and the estimates are nearly equal in others. Although these results demonstrate that using different tests can affect the stability of estimated teacher effects, the cause of those differences is not clear and additional data on how students are prepared for the exams in each district and year and how well the FCAT-NRT aligns with the curriculum might provide insights into the differences.

A decomposition of the variance finds that the sampling error accounts for a smaller portion of the variance in teacher effects estimated with the FCAT-SSS (about 30 percent) than with the FCAT-NRT (about 43 percent) even when using the same students and models. The decompositions also reveal that as with the FCAT-NRT, on average a greater share of the signal variance is within-teacher for elementary school teachers than middle school teachers, although the difference is less pronounced with the FCAT-SSS. Comparisons across models are also similar for the two tests. Estimates for models with partial persistence and student covariate controls have greater reliability than either of the other models. Similarly, the reliability of estimates from models with complete persistence and student fixed effects are again lower than

for other models because of the increase in sampling error, but the differences among models are less pronounced for the FCAT-SSS.

D. Single-Year vs. Multi-year Estimates of Teacher Effects

Even though the year-to-year correlations of estimates of teacher effectiveness derived from models of student achievement fall within the range of those found in other occupations, there may still be concern that value-added estimates of teacher performance may be too variable to be acceptable to stakeholders in a high-stakes accountability system. Given that a large proportion of variance in teacher estimates is due to sampling error, averaging teacher estimates across years is one potential means of significantly reducing inter-temporal variation in measured teacher performance. However, averaging estimates across years can introduce bias if true teacher performance varies across years. This makes averaging across years particularly appealing when employing models with student fixed effects, which may reduce bias from unobserved student heterogeneity but also tend to have larger sampling errors. It also means that averaging estimated effects across years is somewhat less appealing for elementary school teachers, since they appear to exhibit greater within-teacher variation in performance after accounting for sampling errors. Averaging will improve the precision of the estimates but the relatively larger inter-year variability among estimated effects for the same teacher means that the bias due to combining truly different levels of performance may be significant. The mean squared error (MSE), the expected value of the square of the difference between estimated and true performance, is still likely to improve by averaging estimates across years, but bias will offset some of the gains from improved precision and the consequences of biasing the estimates must be considered.

Table 7 presents estimates of the stability coefficient for a single-year and two-year average teacher effect estimates. Consistent with our estimated correlations coefficients, the *Stability* of single-year estimates ranges from about 0.2 to 0.6. In other words, knowing a single year estimate of a teacher's performance reduces our uncertainty in their persistent effect by from 20 to 60 percent depending on the county, model, or grade level. Averaging two years of estimates reduces uncertainty by another 40 to 60 percent, i.e., increases the stability coefficient by 40 to 60 percent. Because persistent effects account for a smaller portion of the variance in single-year estimates for elementary school teachers than middle school teachers, averaging two years is slightly more beneficial for elementary school teachers (on average about a 54 percent additional reduction in prediction error or increase in the stability coefficient for elementary schools compared with a 42 percent reduction for middle schools). However, because of the greater share of variance due to non-persistent change and sampling error in elementary school, the *Stability* of two-year averages is still smaller for elementary school teachers than middle school teachers (on average .45 compared to .56), but the difference between the two groups is smaller than for single year estimates.

One advantage of the stability coefficient is that we can extend it to calculate the stability of teacher effects for three or more years, and easily compare these averages to averages over shorter time spans. For elementary school teachers, averaging three years of estimates would increase the stability coefficient by about 23 percent to roughly .55 on average. For middle school teachers, averaging three years of data only improves that stability by about 18 percent to about .66 on average across models and counties.

V. Summary and Conclusions

While there is keen interest in making personnel decisions based on objective measures of teacher productivity, there is little existing evidence on the inter-temporal variability of teacher effects derived from student test scores. In this paper we construct yearly estimates of teacher productivity from models of student achievement and decompose the variability in those estimates into persistent components, non-persistent changes and sampling errors. We consider the effects of the number of students per teacher, specification of the underlying achievement model, the school district sampled, the test used to measure student achievement, and averaging teacher effect estimates over multiple years.

Consistent with previous research, we find that random variation or sampling error from “noise” in student test scores plays an important role in determining the stability of teacher effect estimates over time. The estimated effects for teachers with only a handful of students in a given year are very imprecise, though the precision increases substantially with the number of students per teacher.

Limiting the samples to teachers with at least 15 students per year and employing achievement models without student fixed effects we obtain year-to-year correlations in estimated teacher effectiveness of 0.22 to 0.67. These correlations imply teacher rankings with only moderate stability; roughly one-third of top-quintile teachers remain in the top quintile the next year while approximately one in ten fall to the bottom quintile of the teacher effectiveness distribution. The results are comparable to those reported by Aaronson, et al. (2007) for Chicago and Koedel and Betts (2007) for San Diego. Though modest, the correlations are also in line with previous research on other occupations, such as insurance salesmen and baseball players, where output is measured directly. Given that bonuses based on productivity are relatively

common for salesmen and professional athletes, it may be that the comparably stable estimates of teacher productivity derived from value-added models could support a performance-based system of compensation for teachers.

Decomposition of the variance in estimated teacher effects underscores the role of sampling error and highlights the inability of currently available measures to explain inter-temporal variation in teacher productivity. We find that approximately one-third to one-half of the variation in teacher effects is simply due to sampling error or “noise” in student achievement. Of the remaining variance, between one-third and two-thirds is attributable to within-teacher variation in effectiveness over time. Consistent with recent literature on teacher training and credentials, very little of the variation in a teacher’s performance over time can be explained by observable teacher characteristics such as experience, attainment of advanced degrees or in-service training. A unique finding of our study is that elementary school teachers appear to have a higher degree of non-persistent change in effectiveness than middle school teachers. This finding warrants future research and may have implications for how pay-for-performance or other systems using value-added measures would function differently in these contexts.

Our decomposition also provides new insights into the specification of achievement models used to generate estimates of teachers’ “value-added.” Using student fixed effects in models of achievement gains, rather than time-invariant student covariates like race and gender, increases sampling error in estimated teacher effects. Student fixed effects control for between-student heterogeneity, but relatively little of the variance in achievement *gains* is between students; rather much of it is within students. Consequently, fixed effects do little to reduce the residual variance in the model. At the same time, the large degrees-of-freedom given to student fixed effects allow them to be collinear with teacher fixed effects, reducing the information

available for estimating teacher effects. The loss of information due to collinearity is greater than the reduction in residual error, resulting in large sampling errors.

The decomposition also reveals that models employing student covariates and assuming partial persistence of prior schooling inputs yield estimates with larger between-teacher variability in the persistent component of teacher effects than other models. By definition, the variability in true teacher effects must be the same for all models. Consequently, larger variance of persistent effects must be due to persistent confounding of the estimated teacher effect by unobserved student attributes due to non-random classroom assignment policies that persist over time (i.e., some teachers always get the “better” students). Using a single prior achievement test score as a proxy for all prior schooling inputs appears to fail to capture some differences among students that gain scores remove. This leads to confounding in estimates that rely on partial persistence that did not result with the other models. The confounding appears to be somewhat greater in middle school, possibly because of tracking of students and teachers consistently teaching the advanced or standard mathematics track across years. This suggests that using lagged achievement as a predictor of current achievement levels might lead to biased estimates of teacher effects. Other authors have come to similar conclusions about this approach (McCaffrey, Han, and Lockwood, forthcoming; Sanders, 2006). Given the limitations of student fixed effects models and the apparent inability of a single lagged test score to account for all prior school-based educational inputs, it may be that we need to relax the assumptions about equal rates of geometric decay across all prior inputs and consider models that include more historical information on students and their teachers, including multiple prior test scores and/or fixed effects for prior teachers.

We also find that the test instrument used to measure student achievement can affect the inter-temporal variability of estimated teacher effects. Using a high-stakes criterion-reference test (the FCAT-SSS) yields different inter-temporal correlations than a low-stakes normed referenced test (the FCAT-NRT) in many cases, though there is no consistent pattern to the differences. However, the decomposition of the variance is generally similar for the two tests when we average across districts, although sampling error generally accounts for a smaller portion of the variance in the estimates from FCAT-SSS which results in greater stability for these estimates compared with estimates based on the FCAT-NRT.

Our decomposition of the variance also allows us to determine the efficacy of employing two-year averages of teacher effects, rather than single-year estimates to determine the relative effectiveness of teachers. Using two-year averages reduces sampling error and increases the ability to predict future teacher performance by roughly 50 percent.

Our findings have important implications for the use of teacher effect estimates in high-stakes teacher retention or compensation decisions. First, one should be very cautious in applying these measures to teachers with few tested students, such as those teaching small classes or large numbers of disabled students who are exempted from standardized tests. Estimates of these teachers' value-added will tend to be over-represented in the extremes of the distribution, so rewarding or penalizing the top or bottom performers would emphasize these teachers and will limit the efficacy of policies designed to identify teachers whose performance is truly exceptional. Second, while averaging teacher performance over multiple years could obscure true changes in teacher performance, there are significant gains in the stability obtained by using two-year average performance measures rather than single-year estimates. Finally, one must recognize that even when multi-year estimates of teacher effectiveness are derived from

samples of teachers with large numbers of students per year, there will still be considerable variability over time. Based on our decomposition analysis and comparisons to other occupations this appears to reflect true changes in teacher performance over time. Nonetheless, adoption of an accountability system based solely on value-added estimates of teacher performance will result in considerable variation in who is rewarded across time.

Although the stability of estimated teacher effects is moderate, they appear to be as stable as estimates of individual worker productivity in other occupations where compensation is a function of performance. However, a policy maker may want to know if the value-added measures are “stable enough” to support policy interventions in the educational context. Of course, this requires knowing the anticipated uses of the measures because different uses will require different levels of stability. One proposed application of estimated teacher effects is in making tenure decisions. The goal of such a program is to improve the average effectiveness of tenured teachers by removing the least effective teachers from the population. The results in the Technical Appendix and our variance decomposition suggest that if a district were to institute a policy where only teachers in the top three quintiles of the distribution of true effectiveness were retained (rather than retaining all teachers), then the average effectiveness of teachers would improve by about 0.04 of a standard deviation unit of student test scores. Applying such a policy with estimated effects rather than true performance would proportionately reduce the gains in the average teacher performance by one minus the square root of the stability coefficient for the estimated effects. With a single year estimate this would imply a reduction of about 47 percent for elementary school teachers and about 38 percent for middle school teachers in the gains to the average performance of tenured teachers afforded by the tenure policy. If the policy were based on a two-year average the gains using the estimated effects would be about one third

smaller for elementary school teachers and about 25 percent smaller for middle school teachers than the gains the policy would have if true effects were observed. Thus, the level of stability in our estimated effects would not lead to excessive losses in the potential for policies to improve the performance of teachers, provided multiple year average effects were used in the policy. However, the very small gains in achievement in average teacher performance that are possible even if the estimates were error free call into question whether such a policy would be likely to lead to substantial gains in student learning.

Given the small variability in persistent effects, even small biases in terms of student variance could have significant effects on the estimates of teacher effects. Also given that errors in the estimates' performance measures depreciate the effects of policies proportional to the square root of the stability which is moderately large for two-year average estimates, bias from omitted variables confounding the teacher effects is likely to remain of greater concern than the inter-temporal instability of the estimates.

Our findings also provide at least some suggestive evidence on how value-added measures might be used in conjunction with other measures of teacher performance in a system of teacher accountability. Given there appears to be substantial variation in teacher performance over time that is not captured by standard measures such as experience and professional development, other, more qualitative measures may serve as a complement in evaluating teachers. For example, Jacob and Lefgren (2008) and Harris and Sass (2007) find that principal ratings of teachers are positively correlated with teachers' ability to boost student test scores. Further, there is currently active research on the development of new classroom observation protocols to measure teacher performance (Blunk, 2007; Danielson and McGreal, 2000; Hill et al., 2007; Pianta and Hamre, forthcoming; Pianta, LaParo, and Hamre, 2006) that may prove

useful and combining these measures with value-added estimates may provide measures of teacher performance with smaller inter-temporal variability.

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." Working Paper #170. Cambridge, MA: Bureau for Research in Economic Analysis of Development (BREAD).
- 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.
- Blunk, M. L. (2007). *The QMI: Results from validation and scale-building*. Chicago: American Educational Association.
- Boardman, Anthony E., and Richard J. Murnane (1979). "Using Panel Data to Improve Estimates of the Determinants of Educational Achievement," *Sociology of Education* 52:113-121.
- Bradbury, John C. (2007). "Does the Baseball Labor Market Properly Value Pitchers?," *Journal of Sports Economics* 8: 616-632.
- Danielson, C. & McGreal, T. L. (2000). Teacher evaluation to enhance professional practice. Princeton, NJ: Educational Testing Service.
- Deadrick, Diana L. and Robert M. Madigan (1990). "Dynamic Criteria Revisited: A Longitudinal Study of Performance Stability and Predictive Validity," *Personnel Psychology* 43: 717-744.
- Goldhaber, Dan and Michael Hansen (2008). "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance." CRPE Working Paper #2008-5.

Gordon, Robert, Thomas J. Kane and Douglas O. Staiger (2006). "Identifying Effective Teachers Using Performance on the Job." White Paper #2006-01. Washington, DC: The Brookings Institution.

Greene, W., *Econometric Analysis*, 4th ed., Prentice Hall, Englewood Cliffs, 2000.

Hanges, Paul J., Benjamin Schneider and Kathryn Niles (1990). "Stability of Performance: An Interactionist Perspective," *Journal of Applied Psychology* 75: 658-667.

Harcourt Assessment (2002). "SAT-10 to SAT-9 Scaled Score to Scaled Score Conversion Tables."

Harris, Douglas N. and Tim R. Sass (2006). "Value-Added Models and the Measurement of Teacher Quality." Unpublished. Tallahassee, FL: Florida State University.

Harris, Douglas N. and Tim R. Sass (2007). "What Makes for a Good Teacher and Who Can Tell?" Unpublished. Tallahassee, FL: Florida State University.

Harris, Douglas N. and Tim R. Sass (2008). "Teacher Training, Teacher Quality and Student Achievement." Unpublished. Tallahassee, FL: Florida State University.

Henry, Rebecca A. and Charles L. Hulin (1987). "Stability of Skilled Performance Across Time: Some Generalizations and Limitations on Utilities," *Journal of Applied Psychology* 72: 457-462.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., et al. (2007). *Mathematical knowledge for teaching and the mathematical quality of instructions: An exploratory study*. Unpublished manuscript.

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

Hoffman, David A., Rick Jacobs and Steve J. Gerras (1992). "Mapping Individual Performance Over Time," *Journal of Applied Psychology* 77: 185-195.

Jacob, Brian A., and Lars Lefgren (2008). "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education" *Journal of Labor Economics* 26: 101-136.

Kane, Thomas J. and Douglas O. Staiger (2002). "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives* 16: 91-114.

Kane, Thomas J., and Douglas O. Staiger (2008). "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working Paper #14607. Cambridge, MA: National Bureau of Economic Research.

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.

Lockwood, J.R., T.A. Louis and Daniel F. McCaffrey (2002). "Uncertainty in Rank Estimation: Implications for Value Added Modeling Accountability Systems," *Journal of Educational and Behavioral Statistics*, 27(3): 255-270.

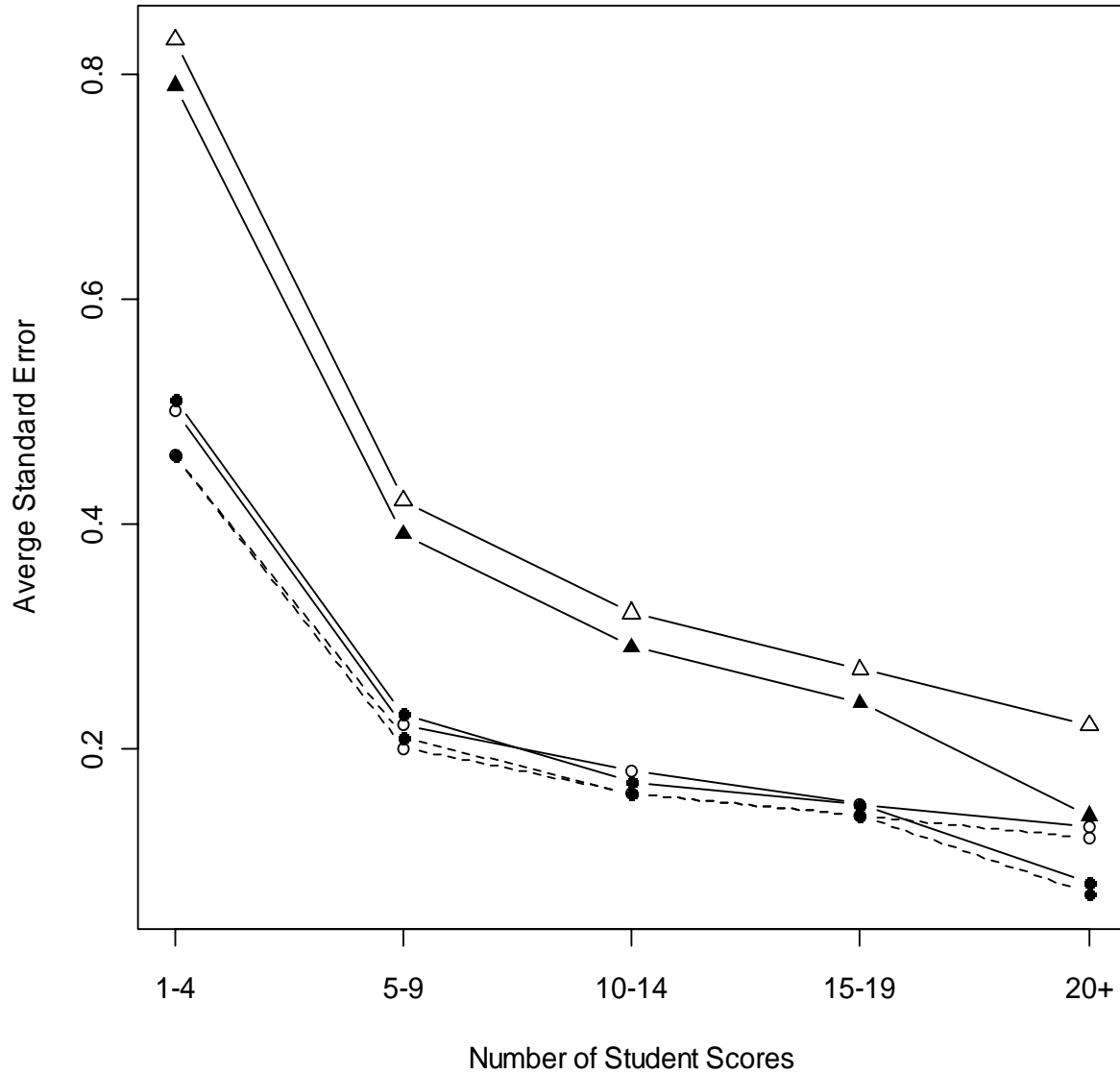
McCaffrey DF, Han B, and Lockwood JR (Forthcoming). "Turning Student Test Scores into Teacher Compensation Systems." In Matthew G. Springer (Ed.) *Performance Incentives: Their Growing Impact on American K-12 Education*. Washington, DC: The Brookings Institute.

Mihaly, Kata, Daniel F. McCaffrey, J.R. Lockwood and Tim R. Sass (2009). "Centering and Reference Groups for Estimates of Teacher Fixed Effects from Value-Added Models," unpublished manuscript.

Pianta, R., & Hamre, B. (forthcoming). Conceptualization, measurement, and improvement of classroom processes: Implications for policy and accountability frameworks. In B. Schneider & G. Sykes (Eds.), *Handbook of Education Policy*.

- Pianta, R. C., LaParo, K. M., & Hamre, B. K. (2006). *Classroom assessment scoring system: Preschool (pre-K) manual*. Baltimore: Brookes Publishing.
- Ployhart, Robert E. and Milton D. Hakel (1998). "The Substantive Nature of Performance Variability: Predicting Interindividual Differences in Intraindividual Performance," *Personnel Psychology* 51:859-901.
- Rambo, William W., Anna M. Chomiak and James M. Price (1983). Consistency of Performance Under Stable Conditions of Work," *Journal of Applied Psychology* 68:78-87.
- Rothe, Harold F. (1978). "Output Rates Among Industrial Employees," *Journal of Applied Psychology* 63: 40-46.
- Rothstein, Jesse (2008). "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." Working Paper #14442. Cambridge, MA: National Bureau of Economic Research.
- Sanders, William L. (2006). "Comparisons Among Various Educational Assessment Value-Added Models" A paper presented at The Power of Two--National Value-Added Conference, Columbus, OH, October 16, 2006.
- Sass, Tim R. (2006). "Charter Schools and Student Achievement in Florida." *Education Finance and Policy* 1(1):91-122.
- Todd, Petra E. and Kenneth I. Wolpin (2003). "On the Specification and Estimation of the Production Function for Cognitive Achievement," *Economic Journal* 113(485):F3-F33.

Figure 1. Mean Standard Error of Estimated Teacher-by-Year Effects by Number of Students per Teacher, Grade Level and Model Type, Pooled Across Five Counties, 2000/01-2004/05



Student controls in the model are denoted by plotting symbols: circles for covariate controls, triangles for fixed effects. Persistence of prior schooling inputs is denoted by line types: solid lines for complete persistence, dashed lines for partial persistence. Elementary grades are denoted by empty plotting symbols and middle grades are denoted by solid plotting symbols.

Table 1. Mean Standard Error of Estimated Teacher-by-Year Effects by Number of Students per Teacher, County, Grade Level and Model Type, 2000/01-2004/05

													Students per Teacher				
			Model Type		1-4		5-9		10-14		15-19		20 or more				
County	Student Controls	Persistence	Mean	No. of	Mean	No. of	Mean	No. of	Mean	No. of	Mean	No. of	Mean	No. of			
			Std. Error	Obs.	Std. Error	Obs.	Std. Error	Obs.	Std. Error	Obs.	Std. Error	Obs.					
													Elementary				
Dade	Covariates	Complete	.48		.22		.17		.14		.12						
	Covariates	Partial	.44	396	.20	240	.15	473	.13	1143	.11	4051					
	Fixed Effects	Complete	.79		.43		.32		.25		.20						
Duval	Covariates	Complete	.51		.22		.17		.15		.13						
	Covariates	Partial	.46	150	.20	210	.16	645	.14	1242	.12	1265					
	Fixed Effects	Complete	.83		.40		.31		.25		.21						
Hillsborough	Covariates	Complete	.55		.22		.17		.15		.13						
	Covariates	Partial	.50	234	.20	157	.16	609	.13	1482	.12	1939					
	Fixed Effects	Complete	.88		.36		.29		.26		.21						
Orange	Covariates	Complete	.47		.23		.18		.16		.14						
	Covariates	Partial	.43	188	.21	419	.17	1186	.15	1617	.13	615					
	Fixed Effects	Complete	.79		.43		.33		.27		.23						
Palm Beach	Covariates	Complete	.55		.23		.18		.15		.13						
	Covariates	Partial	.50	130	.21	140	.16	438	.14	1074	.12	2029					

Students per Teacher

		Model Type		1-4		5-9		10-14		15-19		20 or more	
County	Student Controls	Persistence	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
			Std. Error	No. of Obs.	Std. Error	No. of Obs.	Std. Error	No. of Obs.	Std. Error	No. of Obs.	Std. Error	No. of Obs.	Std. Error
	Fixed Effects	Complete	.92		.51		.36		.29		.25		
Middle													
Dade	Covariates	Complete	.49		.33		.17		.14		.07		
	Covariates	Partial	.44	907	.20	265	.15	137	.13	114	.06	2149	
	Fixed Effects	Complete	.76		.39		.27		.23		.12		
Duval	Covariates	Complete	.51		.23		.19		.15		.08		
	Covariates	Partial	.46	348	.21	85	.17	66	.13	63	.08	981	
	Fixed Effects	Complete	.78		.36		.29		.24		.14		
Hillsborough	Covariates	Complete	.52		.23		.17		.16		.09		
	Covariates	Partial	.47	628	.21	241	.16	151	.14	123	.08	1757	
	Fixed Effects	Complete	.81		.40		.29		.24		.15		
Orange	Covariates	Complete	.51		.23		.17		.15		.08		
	Covariates	Partial	.47	517	.21	171	.16	124	.13	75	.08	1173	
	Fixed Effects	Complete	.81		.38		.29		.25		.15		
Palm Beach	Covariates	Complete	.52		.23		.18		.16		.08		
	Covariates	Partial	.48	550	.21	159	.16	111	.15	107	.07	1243	
	Fixed Effects	Complete	.83		.40		.30		.26		.14		

Table 2. Pooled Year-to-Year Pairwise Correlations of Estimated Teacher-by-Year Effects by County, Grade Level and Model Type, 2000/01-2004/05 (Teachers with ≥ 15 Students in a Year)

County	Model Type		Pooled Year-to-Year Pairwise Correlations
	Student Controls	Persistence	
Elementary			
Dade	Covariates	Complete	.32
	Covariates	Partial	.46
	Fixed Effects	Complete	.23
Duval	Covariates	Complete	.22
	Covariates	Partial	.30
	Fixed Effects	Complete	.26
Hillsborough	Covariates	Complete	.27
	Covariates	Partial	.35
	Fixed Effects	Complete	.24
Orange	Covariates	Complete	.34
	Covariates	Partial	.43
	Fixed Effects	Complete	.39
Palm Beach	Covariates	Complete	.31
	Covariates	Partial	.44
	Fixed Effects	Complete	.16
Middle			

County	Model Type		Pooled Year-to-Year Pairwise Correlations
	Student Controls	Persistence	
Dade	Covariates	Complete	.37
	Covariates	Partial	.67
	Fixed Effects	Complete	.38
Duval	Covariates	Complete	.38
	Covariates	Partial	.53
	Fixed Effects	Complete	.32
Hillsborough	Covariates	Complete	.32
	Covariates	Partial	.47
	Fixed Effects	Complete	.30
Orange	Covariates	Complete	.32
	Covariates	Partial	.61
	Fixed Effects	Complete	.33
Palm Beach	Covariates	Complete	.29
	Covariates	Partial	.51
	Fixed Effects	Complete	.28

Table 3. Variance Decomposition of Estimated Teacher-by-Year Effects by Grade Level and Model Type, 2000/01-2004/05 (Teachers with ≥ 15 Students in a Year)

County	Model Type		Reliability	Proportion of Signal
	Student Controls	Persistence	(Average of the Ratio of Signal Variance to Total Variance) ^a	Variance Due to Variance of Non-persistent Change
Elementary				
Dade	Covariates	Complete	.708	.581
	Covariates	Partial	.764	.422
	Fixed Effects	Complete	.570	.529
Duval	Covariates	Complete	.638	.735
	Covariates	Partial	.700	.603
	Fixed Effects	Complete	.556	.684
Hillsborough	Covariates	Complete	.554	.561
	Covariates	Partial	.654	.490
	Fixed Effects	Complete	.436	.530
Orange	Covariates	Complete	.569	.456
	Covariates	Partial	.673	.354
	Fixed Effects	Complete	.394	.229
Palm Beach	Covariates	Complete	.587	.541
	Covariates	Partial	.736	.409
	Fixed Effects	Complete	.445	.647

County	Model Type		Reliability	Proportion of Signal
	Student Controls	Persistence	(Average of the Ratio of Signal Variance to Total Variance) ^a	Variance Due to Variance of Non-persistent Change
Middle				
Dade	Covariates	Complete	.655	.348
	Covariates	Partial	.820	.256
	Fixed Effects	Complete	.445	.200
Duval	Covariates	Complete	.647	.384
	Covariates	Partial	.774	.293
	Fixed Effects	Complete	.432	.283
Hillsborough	Covariates	Complete	.564	.439
	Covariates	Partial	.713	.396
	Fixed Effects	Complete	.394	.318
Orange	Covariates	Complete	.564	.432
	Covariates	Partial	.799	.270
	Fixed Effects	Complete	.326	.192
Palm Beach	Covariates	Complete	.557	.458
	Covariates	Partial	.799	.348
	Fixed Effects	Complete	.339	.333

^aSignal variance equals the sum of the persistent effect variance and the non-persistent change variance

Table 4. Pooled Quintile Rankings of Estimated Math Teacher Fixed Effects in Year t and Year t+1, 2000/01 - 2004/05: Percent of Teachers by Row (Teachers with ≥ 15 Students in a Year) [Model with Complete Persistence and Student Fixed Effects]

Elementary

		Quintile Ranking in Year t+1				
Quintile Ranking in Year t	County	Bottom 20%	Second 20%	Third 20%	Fourth 20%	Top 20%
Bottom 20%	Dade	30	26	20	14	11
	Duval	33	19	22	14	11
	Hillsborough	33	20	18	14	15
	Orange	41	25	16	9	10
	Palm Beach	31	18	18	18	16
Top 20%	Dade	11	16	16	23	33
	Duval	11	15	14	20	39
	Hillsborough	11	14	16	26	33
	Orange	10	14	18	23	35
	Palm Beach	15	14	20	19	32

Middle

		Quintile Ranking in Year t+1				
Ranking in Year t	County	Bottom	Second	Third	Fourth	Top

		20%	20%	20%	20%	20%
Bottom 20%	Dade	37	21	21	11	10
	Duval	42	22	19	12	5
	Hillsborough	24	28	23	17	8
	Orange	38	27	16	10	9
	Palm Beach	33	21	17	19	10
Top 20%	Dade	7	15	15	28	35
	Duval	8	16	26	18	33
	Hillsborough	8	10	19	26	38
	Orange	13	15	18	26	28
	Palm Beach	9	16	21	23	30

Table 5. Analyses of Individual Performance Over Time in Occupations Other Than K-12 Teaching

Study	Occupation	Output Measure	Frequency of Observations	Period-to-Period Correlation in Output
Rothe (1978)	Four Groups of Foundry Workers	Average quantity of output produced	weekly	.67-.82 (medians)
Rambo, Chomiak and Price (1983)	Two Groups of Textile Workers	Average hourly piece-rate earnings	weekly	.94-.98 (medians)
Deadrick and Madigan (1990)	Sewing Machine Operators	Average hourly piece-rate earnings	weekly	.92 (median, 1-week interval) .55 (23-week interval)
Hoffman, Jacobs and Baratta (1993)	Insurance Salespersons	Value of insurance policies sold	monthly	.22-.63
Ployhart and Hakel (1998)	Securities Analysts	Gross sales commissions	quarterly	.59-.71
Hanges, Schneider and Niles (1990)	University Faculty	Student ratings	semester	.38-.72
Henry and Hulin (1987)	Baseball Hitters	Runs produced	yearly	.47-.59

Study	Occupation	Output Measure	Frequency of Observations	Period-to-Period Correlation in Output
Henry and Hulin (1987)	Baseball Pitchers	Composite of earned run average, walks and strike outs	yearly	.51-.76
Hoffman, Jacobs and Gerras (1992)	Baseball Hitters	Batting average	yearly	.32-.48
Hoffman, Jacobs and Gerras (1992)	Baseball Pitchers	Earned run average	yearly	.12-.45
Bradbury (2007)	Baseball Pitchers	Strikeouts Walks Hit batters Home runs allowed Earned run avg. Batting avg. for balls in play	yearly	.78 .64 .51 .47 .35 .25 (pooled)

Table 6. Pooled Year-to-Year Pairwise Correlations of Estimated Teacher-by-Year Effects by County and Exam Type, 2001/02-2004/05 (Teachers with ≥ 15 Students in a Year)

[Model with Student Fixed Effects and Complete Persistence]

County	Exam	Pooled Year-to-Year Pairwise Correlations
Elementary		
Dade	FCAT-NRT	.17
	FCAT-SSS	.28
Duval	FCAT-NRT	.30
	FCAT-SSS	.16
Hillsborough	FCAT-NRT	.24
	FCAT-SSS	.31
Orange	FCAT-NRT	.41
	FCAT-SSS	.30
Palm Beach	FCAT-NRT	.30
	FCAT-SSS	.48
Middle		
Dade	FCAT-NRT	.40
	FCAT-SSS	.33
Duval	FCAT-NRT	.39
	FCAT-SSS	.39
Hillsborough	FCAT-NRT	.26

County	Exam	Pooled Year-to-Year Pairwise Correlations
	FCAT-SSS	.22
Orange	FCAT-NRT	.41
	FCAT-SSS	.27
Palm Beach	FCAT-NRT	.28
	FCAT-SSS	.38

Note: the sample used in estimating effects includes only observations with both non-missing FCAT-SSS and FCAT-NRT scores.

Table 7. Comparison of Predictive Power of Single-Year and Two-Year-Average Estimated Teacher Effects by County, Grade Level and Model Type (Teachers with ≥ 15 Students in a Year)

			Stability	
Model Type			(Relative Reduction in Prediction Error Variance)	
County	Student Controls	Persistence	Single-Year Estimate	Two-Year Average
Elementary				
Dade	Covariates	Complete	.297	.457
	Covariates	Partial	.442	.612
	Fixed Effects	Complete	.268	.419
Duval	Covariates	Complete	.169	.289
	Covariates	Partial	.278	.435
	Fixed Effects	Complete	.176	.297
Hillsborough	Covariates	Complete	.243	.391
	Covariates	Partial	.333	.499
	Fixed Effects	Complete	.205	.336
Orange	Covariates	Complete	.310	.472
	Covariates	Partial	.435	.606
	Fixed Effects	Complete	.304	.460
Palm Beach	Covariates	Complete	.269	.424

Stability				
Model Type			(Relative Reduction in Prediction Error Variance)	
County	Student Controls	Persistence	Single-Year Estimate	Two-Year Average
	Covariates	Partial	.435	.606
	Fixed Effects	Complete	.157	.268

Middle

	Covariates	Complete	.427	.594
Dade	Covariates	Partial	.610	.755
	Fixed Effects	Complete	.356	.514
	Covariates	Complete	.399	.567
Duval	Covariates	Partial	.547	.705
	Fixed Effects	Complete	.310	.464
	Covariates	Complete	.316	.476
Hillsborough	Covariates	Partial	.430	.599
	Fixed Effects	Complete	.269	.417
	Covariates	Complete	.320	.481
Orange	Covariates	Partial	.583	.735
	Fixed Effects	Complete	.263	.409
	Covariates	Complete	.302	.458
Palm Beach	Covariates	Partial	.520	.681

Model Type		Stability		
		(Relative Reduction in Prediction Error Variance)		
County	Student Controls	Persistence	Single-Year Estimate	Two-Year Average
	Fixed Effects	Complete	.226	.361

Technical Appendix

This technical appendix provides additional details on the relationship between *Stability* and the adjacent year correlation in estimated effects, as well as the relationship between *Stability* and the expected effects of using value-added estimates as the basis for awarding tenure. It also examines the relationship between *Stability* and the properties of quintile rankings.

A. Adjacent Year Correlation

The correlation between estimated effects from two adjacent years is defined as the ratio of the covariance of the estimates to the square root of the product of their variances:

$$\text{Corr}(\delta_{kt}, \delta_{kt+1}) = \frac{\text{Cov}(\delta_{kt}, \delta_{kt+1})}{\sqrt{\text{Var}(\delta_{kt})\text{Var}(\delta_{kt+1})}}.$$

The $\text{Cov}(\delta_{kt}, \delta_{kt+1}) = E((\theta_k + \xi_{kt} + \varepsilon_{kt}) \times (\theta_k + \xi_{kt+1} + \varepsilon_{kt+1})) = E(\theta_k \times \theta_k) = \text{Var}(\theta_k)$, because by definition the sampling errors and non-persistent changes are independent across years and independent of the persistent effects. Assuming the variance of the sampling error and non-persistent change is constant across years then $\text{Var}(\delta_{kt}) = \text{Var}(\delta_{kt+1}) = \tau^2 + \nu^2 + se^2$ and

$$\text{Corr}(\delta_{kt}, \delta_{kt+1}) = \frac{\tau^2}{\tau^2 + \nu^2 + se^2} = \text{Stability}.$$

B. Tenure Awards

We consider an extreme case of the proposal to use value-added estimates in the determination of teacher tenure (Gordon et al., 2006). In this extreme case tenure decisions will be based only on value-added measures and tenure will be awarded to all teachers above the 100th percentile of the distribution of estimated effects. If the persistent effects are truly normally distributed and we could observe persistent effects, then the mean of the distribution of

effects for a tenured teacher is given by the mean of a normal distribution with mean zero and variance τ^2 truncated below at the 100 th percentile:

$$\text{Mean of Persistent Effect of Tenured Teacher Given Perfect Information} = \omega(p)\tau$$

where $\omega(p) = \phi(\alpha_p)/(1-p)$ and ϕ is the standard normal density function and α_p is the p th quantile of the standard normal distribution, i.e., the probability that a standard normal random variable less than or equal to α_p is p (Greene, 2000, p. 900); $\omega(.25) = .42$ and $\omega(.5) = .80$. If we tenured all teachers the mean persistent effect would be zero. Using a cutoff of the 25 th percentile would improve the mean persistent effect among tenured teachers by $.42\tau$ and using the median as the cutoff would improve the mean by $.80\tau$.

If instead the decision was based on estimated effects with total variance $v^2 = \tau^2 + v^2 + se^2$, then the mean persistent effect for tenured teachers would be the expected value of a truncated normal distribution but the truncation is based on a noisy variable. Let θ denote the persistent effect and δ denote the estimated effect. Tenured teachers are then all teachers above the 100 th percentile on the noisy measure; i.e. teachers with $\delta > v\alpha_p$, and the expected persistent effect for these teachers is:

$$\begin{aligned} E(\theta | \delta > v\alpha_p) &= E\{E(\theta | \delta) | \delta > v\alpha_p\} \\ &= E(\tau^2/v^2 \times \delta | \delta > v\alpha_p) \\ &= \tau^2/v^2 \times v\omega(p) \\ &= \sqrt{\tau^2/v^2} \times \tau\omega(p). \end{aligned}$$

Thus, the gains in average teacher performance from using estimated teacher effects for tenure would be $\sqrt{\tau^2/v^2}$, the square root of *Stability*, times as large as the gains from a tenure decision based on the true persistent effects. For stability of roughly .30 we would recover about 55 percent of the maximum gains. If we used a two-year average where we might have stability of about .45 then we would recover about two thirds of the maximum gains from this hypothetical tenure policy. Thus even with the modest reliability of a two-year average of estimated effects, we could recover a substantial portion of the maximum gains.

C. Stability of Quintile Rankings

Assuming a large sample of teachers, normally distributed estimated teacher effects and constant standard errors across teachers and years, then the probability that the estimated effect for a teacher in is in the a quintile in year t and the b quintile in year $t + 1$ is given by the probability of observing values from a bivariate normal distribution in the quintile of the distribution. The probability is given by:

$$P(\delta_{it} \text{ is in quintile } a \text{ and } \delta_{it+1} \text{ is in quintile } b) = \int_{v\alpha_{(2a-1)}}^{v\alpha_{2a}} \int_{v\alpha_{(2b-1)}}^{v\alpha_{2b}} \phi\left(\frac{\delta_t}{v}, \frac{\delta_{t+1}}{v} \mid S\right) d\delta_t d\delta_{t+1}, \quad (A1)$$

assuming that $\text{Var}(\delta_{it}) = \text{Var}(\delta_{it+1}) = v^2$, $\Phi(\alpha_{.2a}) = .2a$ for $a=0, 1, \dots, 5$, where $\Phi(u)$ denotes the cumulative density function of the standard normal distribution evaluated at u and $\phi(u_t, u_{t+1} \mid \rho)$ denotes a bivariate normal density with zero means, variances one and correlation ρ . As shown above, the correlation of the two estimated effects is the *Stability*, S . Equation A1 yields,

$$P(\delta_{it+1} \text{ is in quintile } b \mid \delta_{it} \text{ is in quintile } a) =$$

$$\frac{\int_{v\alpha_{(2a-1)}}^{v\alpha_{2a}} \Phi\left(\frac{v\alpha_{.2b} - S\delta_t}{v\sqrt{1-S^2}}\right) \phi\left(\frac{\delta_t}{v}\right) d\delta_t - \int_{v\alpha_{(2a-1)}}^{v\alpha_{2a}} \Phi\left(\frac{v\alpha_{.2(b-1)} - S\delta_t}{v\sqrt{1-S^2}}\right) \phi\left(\frac{\delta_t}{v}\right) d\delta_t}{.2}. \quad (\text{A2})$$

It is clear from Equation A2 that the conditional probabilities are a function of *Stability*. We evaluated the equation via Monte Carlo simulation and plot values as a function of *Stability* for $a = 1$ and $b=1, 2, 3, 4,$ and 5 in Figure A1.

Figure A1. Probability of a Teacher's Estimated Effect Being in Quintile q in Year $t+1$ Given the Estimated Effect Was in Quintile 1 in Year t as a Function of *Stability*

