THE INTERTEMPORAL STABILITY OF TEACHER EFFECT ESTIMATES *

by

Daniel F. McCaffrey
The RAND Corporation

Tim R. Sass
Florida State University

J. R. Lockwood
The RAND Corporation

Original Version: April 9, 2008
This Version: June 27, 2008

Abstract

Recently, a number of school districts have begun using measures of teachers' contributions to student test scores or teacher "value added" to determine salaries and other monetary rewards. In this paper we investigate the precision of value-added measures by analyzing their inter-temporal stability. We find that these measures of teacher productivity are only moderately stable over time, with year-to-year correlations in the range of 0.2-0.3. However, dis-attenuated year-to-year correlations are much higher, suggesting that much of the variation in measured teacher performance is due to random error or "noise" in the average test score gains of a teacher's students. We also find that changes to the specification of the achievement model used to generate teacher effects generally have little impact on the stability of the resulting value-added measures. The one exception being when student covariates are used to represent student heterogeneity rather than student fixed effects; in some settings this resulted in a substantial increase in the cross-year correlation. This indicates there may be non-random assignment of students to teachers based on unobserved student characteristics that can affect the stability of teacher effect estimates. Finally, we re-estimate the achievement model using an alternative test score measure. The observed variation in measured teacher performance in some cases changes significantly across tests, implying that changes in the test instrument over time can affect variability in measured teacher effectiveness as well.

---

*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.

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. Any errors in estimated effects, systematic or otherwise, can result in rewarding or retaining teachers who are not truly the most effective.

Proposals to use measured early-career performance in tenure decisions and to adjust teachers' base salaries according to their current measured performance are predicated on the tacit assumption that a teacher's current performance is a good indicator of his or her future performance in the absence of an intervention of formal or informal training. Consequently it is important to determine how errors in estimated teacher effects might contribute to inter-temporal instability of effects and potentially undermine policies meant to improve the teacher labor force. It is also important to test the tacit assumption of stability in teacher performance and identify sources of year-to-year variability in teacher performance in order to develop appropriate policies toward teacher compensation, retention and professional training.

Only a few previous studies have measured the stability of teacher effects and none have analyzed the determinants of inter-temporal stability in any 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 than 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 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 stay in the same quintile from one year to the next, the degree of persistence is less than that found by Aaronson, et al. 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 which include student, teacher and school fixed effects). Omitting student and school fixed effects they find the teacher effects to be more stable over time; 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 are also in the top quintile in the second year. The enhanced stability could simply be due to changing from a within-school measure of teacher performance (when school effects are included) to an across-school measure (when school effects are omitted). Alternatively, the apparent increase in intertemporal stability could be an indication there exists a persistent non-random assignment of students to teachers. Omission of the student fixed effects, which control for unobserved student heterogeneity, could produce estimated teacher effects that are both biased and more stable than the true teacher effects.

In this paper we consider the inter-temporal stability of teacher performance measures based on a teacher's estimated contribution to student achievement. We begin by examining the inter-temporal variation in the average achievement gain of a teacher's students relative to the students' long-run or baseline performance, using a simple student-fixed effects model to estimate teacher performance. Both year-to-year correlations in the estimated teacher effects and year-to-year quintile rankings of teachers' estimated value-added indicate only moderate inter-temporal stability in measured teacher performance. However, dis-attenuated year-to-year correlations, which estimate what the correlation of effects would be had they been estimated without random errors, are much higher, suggesting that much of the variation in observed performance is due to random error or "noise" in average test score gains of a teacher's students. In order to better understand the sources of inter-temporal instability we decompose the variation in measured teacher performance into variation across teachers and within teachers over time.

The within-teacher variation is further divided up into the proportions that can be explained by variation in students and their peers, teachers and schools over time. We find that the variance of teacher effects (adjusted for random errors) is greater in elementary schools than in middle schools, that the proportion this variance that is within-teachers also tends to be greater in elementary than in middle school and that the greater the within-teacher variation, the smaller the proportion that can be explained by variation in observed student/peer, teacher and school characteristics. Next, we re-estimate the achievement model used to generate teacher effects, employing a variety of controls for student, peer, teacher and school characteristics. In most cases the model specification has little effect on the stability of teacher effects. The one exception being when student covariates are used to represent student heterogeneity rather than student fixed effects. This suggests there may be non-random assignment of students to teachers based on unobserved student characteristics that can affect the stability of teacher effect estimates. Finally, we re-estimate the achievement model using an alternative test score measure. The observed variation in measured teacher performance in some cases changes significantly across tests, suggesting that changes in the test instrument over time can affect variability in measured teacher effectiveness as well.

II. Methods

A. The Components of Student Achievement Gains

Following Rivkin, Hanushek and Kain (2005) one can decompose the achievement gain of student *i* in classroom *j* taught by teacher *k* in school *m* at time *t* into a set of fixed and time-varying components:

$$\Delta A_{ijkmt} = \alpha_i + \delta_k + \theta_m + \nu_{ijkmt} \quad (1)$$

The test score gain is assumed to be an additively separable function of student (α), teacher (δ) and school (θ) fixed effects plus an error that is a composite of time-varying factors.[1] Equation (1) does not include a time invariant classroom component because students are grouped into different classrooms in each time period and consequently classrooms do not persist across periods. This model is based on several assumptions about the inputs into cumulative student achievement. In particular, it requires that a teacher's effect on a student's achievement level does not decay over time. Additional details on the assumptions necessary to derive equation (1) from a more general cumulative achievement function are found in Boardman and Murnane (1979), Todd and Wolpin (2003) and Harris and Sass (2006).

If we decompose the time-varying component into student (γ_{it}), classroom peer (π_{jt}), teacher (ν_{kt}) and school (ψ_{mt}) factors plus random error (ε_{ijkmt}) and assume additive separability among the time-varying factors we obtain:

$$\Delta A_{ijkmt} = (\alpha_i + \gamma_{it}) + \pi_{jt} + (\delta_k + \nu_{kt}) + (\theta_m + \psi_{mt}) + \varepsilon_{ijkmt} \quad (2)$$

A teacher's contribution to student learning at any given time *t* is the sum of the teacher fixed and variable components, ($\delta_k + \nu_{kt}$). We refer to this as the time-varying teacher effect, ϕ_{kt}. The time-varying teacher effect can be expressed as the difference between the average achievement gain for all students taught by teacher k at time t minus the group averages of the other determinants of achievement gains:

$$\phi_{kt} = \overline{\Delta A}_{kt} - (\overline{\alpha}_{kt} + \overline{\gamma}_{kt}) - \overline{\pi}_{kt} - (\overline{\theta}_m + \overline{\psi}_{mt}) - \overline{\varepsilon}_{kt}\,, \quad (3)$$

[1] The assumption of additive separability assumes that teacher productivity is independent of student characteristics and the school environment.

where $\overline{\Delta A}_{kt}$, $\bar{\delta}_{kt}$, $\bar{\pi}_{kt}$, $\bar{\theta}_{m}$, $\bar{\psi}_{mt}$, and $\bar{\varepsilon}_{kt}$ denote the average values of the corresponding variables averaged over all students taught by teacher k in time t.

Of course the true achievement gains of students as well as the attributes of students, classroom peers and schools that determine achievement gains are never fully observable. The availability of data on observable time-varying student, peer and school characteristics will vary across data sets. Even in the best administrative databases the number of exogenous time-varying student, peer and school characteristics is typically quite limited. In order to see how data availability and other sources of error contribute to variation in time-varying teacher effects we begin with a scenario where no time-varying teacher, peer or school components are observable. We also exclude school fixed effects in order to allow for comparison of teachers across schools. Thus the time-varying teacher effect becomes:

$$\phi_{kt} = \overline{\Delta A}_{kt} - \bar{\alpha}_{kt} + \eta_{kt} \tag{4}$$

where $\eta_{kt} = -\bar{\gamma}_{kt} - \bar{\pi}_{kt} - \bar{\theta}_{m} - \bar{\psi}_{mt} - \bar{\varepsilon}_{kt}$

The estimated time-varying effect for teacher k at time t, $\hat{\phi}_{kt}$, equals the average student gains for teacher k in time t relative to those students' baseline gains or "fixed effects", $\overline{\Delta A}_{kt} - \bar{\alpha}_{kt}$. To study the inter-temporal variability in estimated teacher effects we can decompose $\hat{\phi}_{kt}$ into its components:

$$\begin{aligned} \hat{\phi}_{kt} &= \phi_{kt} - \eta_{kt} \\ &= \delta_{k} + \nu_{kt} + \bar{\gamma}_{kt} + \bar{\pi}_{kt} + \bar{\theta}_{m} + \bar{\psi}_{mt} + \bar{\varepsilon}_{kt}. \end{aligned} \tag{5}$$

Hence the sources of variance in estimated effects are: time-invariant teacher attributes, time-varying teacher variables, aggregates of time-varying student variables, time-varying peer variables, time-invariant school variables, time-varying school variables and random errors.

Variability in each of the components contributes to variability in estimated effects among teachers and to year-to-year variability of estimates within teachers to the extent that each component varies among and within teachers. The variability in the non-teacher components are of two types: individual student-level errors which are independent across students and errors that are correlated across students within a teacher. The later could result from individual-level student shocks that are correlated within a teacher or from more aggregate classroom-level or school-level shocks that impact all students within a classroom or within a teacher in a given year.

Errors that are independent across students could include things like whether a student had a good sleep the night before the exam or whether they are bothered by personal issues on the day of the exam. They also include variability due to potentially observable student characteristics which are not accounted for by the model, but only to the extent these variables vary like they would if students were randomly assigned. Any excess heterogeneity at the teacher level is not included in this source of variability.

These independent student-level errors are the only component of student error accounted for by the standard errors of the estimated teacher effects because, almost universally, estimation methods assume independent random errors. Methods that account for the standard errors when making inferences about teachers (e.g., confidence intervals, significance tests, and empirical Bayes shrinkage) will mitigate the contributions of independent student-level errors to inferences about teachers. Moreover, independent student-level errors are by definition uncorrelated with

classroom assignment and do not contribute to bias in the estimated teacher effects. The variability in these errors will decrease with the number of students used to estimate a teacher's effect.

Aggregate or teacher-level errors are individual-level errors that are correlated across students in the same class or errors from more aggregate-level shocks to classrooms or schools. Unlike independent student-level errors, errors that are common across students within a teacher are not accounted for by the standard errors of estimated teacher effects since the standard errors are estimated under the assumption that random errors in equation (5) are independent. Consequently they are a potential source of inter-temporal variance that affect year-to-year correlations among estimated teacher effects even after adjustments are made for attenuation of the correlation because of independent student-level errors.

Errors that are correlated across students within a teacher can be caused by either random shocks or by non-random selection. Random shocks could be student-level errors that are correlated within a teacher. For example, given physical proximity, the likelihood a child is ill when taking the exam may be correlated across students within a class. Alternatively, random classroom-level shocks, such as the classroom's air conditioning failing on exam day or the oft-imagined barking dog outside the classroom would produce a common error among students in a classroom. Likewise, school-level random shocks, like a car crash on the street beside the school or the school's entire heating system malfunctioning on exam day would create a common error within a teacher.

These correlated or common shocks are one-time events that occur during a single year of testing and which do not persist overtime. Because they simultaneously affect multiple students they are not accounted for by the estimated standard errors of teacher effects. However, because

they do not persist across years and are by definition unrelated to student background variables, these shocks contribute to inter-temporal variability but not to bias. In general, increasing the number of students per classroom will not reduce the contribution of these shocks to inter-temporal variation in measured teacher performance. However, if the shocks are at the classroom level (and not at the school level), then teachers who teach multiple classes per year (eg. middle school teachers) would experience less variability in their measured performance than would teachers who teach a single class (e.g., elementary school teachers).

Correlated or teacher-level errors can also result from the non-random assignment of students to teachers or teachers to schools. For example, consider the matching of students to teachers in classrooms. In any given year, some teachers might be consciously assigned relatively well-behaved students, which leads to fewer classroom disruptions, more learning time and higher average test score gains. To the extent that student behavior is not observed and taken into account, the estimated teacher effects for the teachers with well-behaved students would be greater than the teacher's true effect.

The effect of this non-random error on the stability of measured teacher effects depends on the degree of inter-temporal variation in the underlying scheme of assigning students to teachers. If the assignment rule is constant over time (i.e., the same teachers get the best-behaved students year in and year out) then the teacher effects would appear to be more stable then they actually are due to the stability in the underlying student behavior that is being falsely attributed to the teacher. In contrast, if the assignment rule changes over time, this could decrease the inter-temporal stability of measured teacher effects. For example, suppose that some principals tend to assign the most unruly students to the most proficient teachers whereas other principals do the opposite and saddle the least productive teachers with the most disruptive

students. A high degree of principal turnover would lead to large swings in the behavior of students a teacher faces from year to year and consequently (to the extent that student behavior is not explicitly accounted for) produce large within-teacher variation in measured effectiveness over time. The contribution of errors from selection to inter-temporal variability could be large even for teachers with large classes because the variance of these errors does not decrease with increases in class size.

Beyond their contributions to inter-temporal instability, time-varying errors from selection, like time-invariant errors from selection, contribute to systematic errors in estimated teacher effects in which the differences between the estimated and true effects are correlated with student or school characteristics even after accounting for the standard errors of the estimated effects. By falsely attributing the effects of unmeasured student or school characteristics to teachers, such systematic errors and the resulting fallacious inference about teachers could be problematic for any system that uses value-added performance measures for high stakes decisions. Stakeholders could rightly argue that rewards are determined by assignments of teachers to schools and assignments of teachers to classrooms within schools, rather than to true differences in teacher productivity. Thus errors from selection could limit the utility of estimated teacher effects even beyond any problems caused by their potentially large contribution to inter-temporal variability in estimated effects.

The decomposition in equation (5) is motivated by models (1) to (4) and the fixed effects estimators used in this paper. However, the decomposition is generic and the sources of error we identify are the sources that will contribute to the variability in estimated teacher effects generated using any estimation method. The relative contributions of the various components will depend on the data and estimation method but the decomposition applies generally.

The decomposition also applies when the model used to generate estimated teacher effects is misspecified. For example, misspecification might result if student gains are not additively separable because teacher effects depend on the students' characteristics or because of unique interaction between teachers and their classes. Misspecification also might result if the assumptions about the persistence of the inputs used to derive model (1) are incorrect. The misspecification would contribute additional errors in the estimated effects and these errors would also contribute to inter-temporal variability of estimated effects. However, the contributions of errors due to misspecification could be decomposed into the same components used in Equation 5, so that the decomposition still holds. Misspecification could change the potential contributions of various sources, however. For example, if the model is misspecified, then observable student variables might contribute to inter-temporal variability even if they are used in the estimation process. Consequently, studying the contributions of various components to inter-temporal variability might be useful even if a very rich model were used in the estimation.

B. Additional Value-Added Models of Teacher Quality

Equation 4 presents a model with no observable measures on students, peers, or schools. We use that model to generate estimated effects and decompose the potential sources of error and inter-temporal variance in estimated teacher effects. As just discussed, the decomposition identifies the sources of errors from omitted variables and potential misspecification. To understand the effects of omitted variables, we also fit a series of models that include additional factors and calculate correlations between the resulting estimated effects from adjacent years. These correlations provide a means of interpreting the decomposition and provide direct estimates of stability in estimated effects from alternative models.

The additional value-added specifications we explore are variants of the following model:

$$\Delta A_{it} = \boldsymbol{\beta}_1 \mathbf{X}_{it} + \boldsymbol{\beta}_2 \mathbf{P}_{-ijt} + \boldsymbol{\beta}_3 \mathbf{T}_{kt} + \boldsymbol{\beta}_4 \mathbf{S}_{mt} + \alpha_i + \phi_{kt} + \mathrm{e}_{ijkmt} \quad (6)$$

which includes time varying student/family inputs, $\mathbf{X}_{it}$, classroom peer characteristics, $\mathbf{P}_{\text{-ijmt}}$ where the subscript –i denotes students other than individual i in classroom j in school m, and school-level input, $\mathbf{S}_{\text{mt}}$, unmeasured time-invariant student/family characteristics, represented by a student fixed effect, α_i, time-varying teacher characteristics captured by a year-specific teacher effect, ϕ_{kt}, and a random error term. Different specifications include various permutations of the student, school and peer variables.

C. Estimation of Teacher Effects

To estimate equations (4) and (6) we use standard fixed-effects regression techniques, including fixed effects for both students and teachers along with additional explanatory variables as determined by our particular specification. Because the data include test scores from students at multiple grade levels each year, the model also includes separate means by grade level by year.

In addition to the standard teacher fixed effect estimates, we also generate Empirical Bayes (EB) shrunken estimates of the teacher effects. Given we are making many estimates of similar quantities (i.e. estimating effects for many teachers, $\hat{\phi}_{kt}$), the accuracy of the estimates can be improved using the estimates $\tilde{\phi}_{kt} = B_{kt}\hat{\phi}_{kt}$, rather than $\hat{\phi}_{kt}$ (Morris, 1983). The factor $B_{kt} = A/(A + se_{kt}^2)$, where A equals the variability of the true values of the teacher effects, the ϕ_{kt}. Because B_{kt} is less than or equal to one, $\tilde{\phi}_{kt}$ is shrunken back toward zero relative to $\hat{\phi}_{kt}$ and these estimates are referred to as shrunken estimates. The estimates are often referred to as Empirical

Bayes estimates because they can be motivated as solutions to a Bayesian estimation process with some prior distributions replaced by parameters estimated from the data. The challenge in creating EB estimates is estimating *A*. A method of moments approach is commonly used (Morris, 1983) as is maximum likelihood (Carlin and Louis (2000)). We use maximum likelihood estimates described below in Section D. We produce EB estimates for every teacher from each district each year and study the stability of these estimates along with the raw estimates.

Several inter-related complications arise when estimating annual fixed effects for a large sample of teachers. The primary problem arises because the estimation method cannot uniquely estimate an effect for every teacher while simultaneously controlling for fixed effects for all the individual students. The inability to uniquely estimate every parameter is known as a lack of identification of the teacher effects. Provided no other complications arise, the lack of identification is no different than the lack of identification for estimating effects for any categorical variable using linear regression models.

The standard solution for the lack of identification is to estimate contrasts of teacher effects rather than individual teacher effects. For example, the default solution taken by 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. An alternative is to contrast all teachers to the average teacher effect. As we describe below we believe this alternative is strongly preferable to contrasting teachers to an arbitrary holdout teacher.

However, features of our test scores data require additional restrictions for identification of teacher effects. In some years elementary and middle school grade teachers are disjoint in that there are no teachers who teach both elementary and middle school students. When the groups

are disjoint, we cannot uniquely determine grade-level means by year and teacher effects by year. To identify teacher effects we must use contrasts of teacher effects among teachers in the same grade-level group, e.g., among elementary school teachers or among middle school teachers. Hence we define the teacher effect to be the difference between the mean for an individual teacher and the mean for all teachers in the same grade-level group (elementary or middle) for a given year.

A second feature of the data that results in additional constraints on our estimated effects is stratification (McCaffrey et al. (2004)) or lack of connectedness (Searle (1971)). Stratification occurs when estimating models with both student and teacher fixed effects if there are groups of students and teachers in which none of the students in one group ever shares a teacher with any of the students in the other groups. This problem has also been identified in the economics literature on the estimation of employee and firm effects (Cornelißen (2006), Abowd et al (2002)). However, in that literature, firm effects are assumed to persist across years, whereas in our problem we wish to estimate separate teacher effects each year.

When stratification occurs, unique teacher-effect contrasts can only be estimated relative to the other teachers in each stratum. Because teachers in different strata do not share any common students we cannot distinguish between differences among the average student across strata and differences among the average teacher effects across strata. Hence we estimate effects that equal the difference between a teacher's effect and the average of all teachers in his or her stratum and grade-level group (elementary or middle school) by year.

There are several advantages to estimating the effects as we have parameterized them. Given the constraints of stratification and grade-level groupings of teachers, the estimated effects must be relative to strata and grade-level groupings. Defining the estimates relative to the mean

of the teachers within a grade-level group has substantive justifications. Under the appropriate assumptions, these estimates can be interpreted as causal effects of a teacher in the sense that they are the difference between the student's potential outcomes when taught by their observed teacher and their potential outcome when taught by the average teacher in the stratum and group. Given that students will not change grade-level groupings within a year, restricting the comparison to teachers within the same grade-level group seems appropriate. Similarly, from a substantive standpoint, comparing a teachers' performance only to other teachers teaching the same grade level helps ensure that we compare teachers performing similar functions in similar environments. Even if we want to compare across grade-level groups, it seems reasonable to do so using teachers' performance relative to their peers teaching in the same grade-level group.

There is no obvious substantive justification for contrasting teachers to the strata means but this choice can make estimates invariant to arbitrary decisions on which teacher to use as a holdout and removes certain types of inter-temporal instability from estimated effects. Alternatively, contrasting teachers to different arbitrary holdout teachers each year could lead to considerable inter-temporal variability in any given teachers' estimated effects. Provided we constrain our estimates to within strata, this inter-temporal variability would not affect a teacher's relative position in the distribution of estimated performances. It would not affect rankings and cross-year correlations, nor would it add to instability in bonuses based on relative position. However, teachers' raw estimates could vary greatly from year and this could cause teachers great confusion. Post hoc centering of effects could remove this source of inter-temporal variability but then the estimated standard error of the estimates would be wrong and computationally challenging to correct with large samples of teachers. If data are pooled across

strata, then the choice of an arbitrary holdout would affect a teacher's relative position in the distribution of teacher effect estimates.

Contrasting teachers to the average teacher within the strata might not completely eliminate these problems, but should greatly mitigate them. For every teacher, changes in the average of the performance of the other teachers in his or her stratum will change the estimate of his or her performance in any given year. However, in the data used in this study, in each county, year, and grade-level, there is one large stratum with well over 90 percent of the teachers and students and a few small strata with very small numbers of teachers and students in each one. Hence, year-to-year variation in the strata means should be inconsequential for nearly all teachers. By contrasting teachers to the strata means, both relative performance and raw estimates of performance will not be influenced by the selection of an arbitrary holdout teacher and the standard errors will be correct without complex additional computations.

Our parameterization has even greater advantages if one wishes to use Empirical Bayes shrinkage post hoc to improve the precision of estimates. EB estimates shrink the teacher effects back to zero. When effects are relative to an arbitrary holdout teacher (as is the case in most popular statistical software packages), EB estimates shrink each effect to the effect of the holdout teacher, which could be any value. The rank ordering of teachers, cross-year correlations, and inter-temporal stability of estimates can all be sensitive to the chosen holdout. Furthermore simple post hoc re-centering of EB estimates will not remove the inter-temporal instability introduced by using an arbitrary holdout teacher. Re-centering would need to occur before EB shrinking, but this would make the standard errors and the resulting shrinkage factors incorrect, unless cumbersome and computationally intensive adjustments were made to the

standard error. These problems are all avoided by our chosen parameterization of teacher effects.

D. Estimating Stability and Sources of Variance

As in previous research (Ballou (2005), Aaronson et al. (2007), Koedel and Betts (2007)), we use the correlation between estimated teacher effects from adjacent years and the stability of teacher rankings to describe the stability of estimated teacher effects. The value of these measures is they provide a summary of how stable raw annual teacher effect estimates are likely to be and provide guidance to researchers and policy makers on the potential limitations of such estimates. The disadvantage of these measures of stability is that they provide no information about the sources of instability.

We conduct two additional analyses to unpack the sources of the instability. The first explores how much independent student errors contribute to instability. We model each estimated teacher effect as the sum of two components: $\hat{\phi}_{kt} = \chi_{kt} + \xi_{kt}$, where ξ_{kt} is the error due to independent student errors, with variance equal to the standard error squared, and χ_{kt} contains all other sources of variance in the estimated effect. Given this decomposition, we assume that the estimates for teacher *k* in years *t* and *t+1*, $\left(\hat{\phi}_{kt}, \hat{\phi}_{kt+1}\right) \sim N(\boldsymbol{\mu}, \boldsymbol{V})$, where $\boldsymbol{\mu}$ is a vector means for years *t* and *t* + 1, and $\mathbf{V} = \boldsymbol{\Sigma} + \boldsymbol{\Omega}$, with $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{tt} & \sigma_{tt+1} \\ \sigma_{tt+1} & \sigma_{t+1t+1} \end{pmatrix}$, $\boldsymbol{\Omega} = \begin{pmatrix} se_{kt}^2 & 0 \\ 0 & se_{kt+1}^2 \end{pmatrix}$, and se_{kt}^2, se_{kt+1}^2 equal to the squared standard errors of the estimated teacher effects provided by the statistical software under the assumption of independent residual errors among students from the same classroom. We estimated the components of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ using maximum likelihood methods to calculate $r = \hat{\sigma}_{tt+1} \Big/ \sqrt{\hat{\sigma}_{tt}\hat{\sigma}_{t+1t+1}}$, which equals the correlation of the effects had it not been

attenuated due to the presence of independent student-level errors. We call *r* the dis-attenuated correlation.

To estimate the sources of variance presented in Equation 5, we first decompose each estimated teacher effect into $\hat{\phi}_{kt} = \chi_k + \chi_{kt} + \xi_{kt}$, where again ξ_{kt} is the error due to independent student-level errors and we now distinguish between time-varying sources of variability, $\chi_{kt.}$, including classroom level-shocks, and time-varying errors from selection for students and schools, and variation due to time varying teacher variables or other year-to-year variation in true teacher performance, and time-invariant sources of variability among teachers, χ_k, including time-invariant teacher attributes and performance, and time-invariant errors from selection of students or schools. We assumed that $\chi_k \sim N(0, \tau^2)$ and $\chi_{kt} \sim N(\mu_t, \kappa^2)$ and ξ_{kt} has variance equal to the square of the standard error for the estimated effect. These assumptions yield a joint normal likelihood function for the estimated effects and we maximize this function to develop estimates of τ^2 and κ^2.[2]

We explore the contribution of classroom averages of observed time varying student variables, $\boldsymbol{X}_{kt}$, and time-varying teacher characteristics, $\boldsymbol{Z}_{kt}$, to this component of variability. Using variables described below we fit the linear model $\hat{\phi}_{kt} = a_k + \boldsymbol{X'}_{kt}\boldsymbol{\beta} + \boldsymbol{Z'}_{kt}\boldsymbol{\gamma} + e_{kt}$ using fixed effects for individual teachers and weighting the observations to account for the heteroskedasticity in e_{kt}. We then calculate the within-teacher variance of $z_{kt} = \boldsymbol{Z}'_{kt}\hat{\boldsymbol{\gamma}}$, $\hat{\sigma}_z^2$. We also calculate $x_{ktj} = \boldsymbol{X}'_{ktj}\hat{\boldsymbol{\beta}}$ for every student in every teachers' classroom for every year. We

[2] To improve estimation we estimated the natural log of τ^2 and the natural log of the ICC/(1-ICC), where ICC= $\tau^2/(\tau^2 + \kappa^2)$.

decompose the variance in x_{ktj} into the within-classroom-and-year component which is independent student-level error and a component of the standard errors, the within teacher-between year component, $\hat{\sigma}_x^2$, which is a potential source of time-varying error due to selection and contributes to κ^2, and the between-teacher component, which is a source of time-invariant error due to selection and contributes τ^2. We calculate the ratio of $\hat{\sigma}_z^2$ and $\hat{\sigma}_x^2$ to our estimate of κ^2 to determine the share of the inter-temporal variance explained by observed teacher and observed time-varying student characteristics.

III. Data

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.[3] From this statewide database we select data from four large school districts in the state, Duval, Hillsborough, Orange and Palm Beach. Each of the four 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 the data from the entire state, selection of these four 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

[3] Detailed descriptions of the Florida data are provided in Sass (2006) and Harris and Sass (2008b).

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.[4] 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.

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 school grades and make student retention decisions. In our application the scores are normed to have mean zero and standard deviation one each grade and year. 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). The scores on the Stanford-9 are scaled to a single developmental scale so that a one-point increase in the score at one place on the scale is meant to be equivalent to a one-point increase anywhere else on the scale. The Stanford-9 is a vertically scaled exam, thus scale scores typically increase with the grade level. We rely primarily on the FCAT-NRT exam since it

[4] 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.

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.

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, we restrict our analysis of student achievement to elementary students in "self-contained" classrooms and middle-school students who are taking only one math course. We also exclude students who are repeating a grade. However, all students enrolled in a course are included in the measurement of peer-group characteristics used in our analyses of the sources of year-to-year variance in estimated effects. 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 time-varying observable characteristics of students, peers, teachers and schools. At the student level we observe 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" (where fewer than 30 percent of students in the same initial school and grade made the same move). 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 the previous two years), educational attainment (captured by an indicator for possession of an advanced degree), and an indicator of whether or they are fully certified or hold a temporary license. At the school level with 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. 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.

To ensure comparability across analyses, including those that do and do not involve the observed time-varying variables, we restrict the sample to only those student observations with non-missing data on all of the student, peer, teacher and school variables. This ensures that both our student and teacher samples are the same for all analyses. For analyses comparing the FCAT-SSS and FCAT-NRT exams (Table 6) we exclude any observations that lack valid data on both exam scores, ensuring comparability in the estimation samples.

IV. Results

A. Baseline Model

Table 1 provides estimated cross-year correlations of the time-varying teacher effect with no controls other than student fixed effects (equation (4)). Student achievement gains are measured by the annual change in "normed" FCAT-NRT scores (i.e., prior to calculating gain scores the FCAT-NRT scores were linearly transformed to have mean zero and standard deviation one for each grade level for each year). The estimation of teacher effects is repeated for each of our four districts, with separate effects estimated for each teacher in each year. However, all years of data within a district are analyzed simultaneously to make the most efficient use of the longitudinal achievement data for individual students. Since fourth graders in 2004/05 are only observed for a single year, they have only one observed achievement gain and thus drop out of the model with student fixed effects. Consequently, we do not report the correlation in estimated teacher effects at the elementary level for the final pair of years.

The correlations are generally moderately low to moderate, mostly in the range of 0.2 to 0.3. There are no obvious patterns across time or across school districts. Table 1 also presents dis-attenuated correlations, which are generally much higher, usually falling between 0.5 and 0.8, suggesting that independent student-level variation is a significant source of instability in estimated teacher effects. In some cases the dis-attenuated correlations approach a value of one. Typically this occurs in situations where the variance from all sources other than independent student-level variation is very small, thereby producing general instability in the estimated dis-attenuated correlation. Very high values might also indicate that some of the assumptions used in estimating the dis-attenuated correlations may be incorrect. However, the qualitative findings are clear: independent student-level variation is a significant source of the variability in teacher

effects; and in some settings there is very little evidence of variance in estimated effects among or within teachers beyond that caused by these random individual-level shocks.

Table 1 also presents the cross-year correlation of the estimated effects that were adjusted by EB shrinkage. Although EB shrinkage increases the precision of the estimates so that any given teacher's effects will vary less from year to year, it has a relatively limited effect on the cross-year correlation. For middle school teachers, EB shrinkage consistently improves the correlation whereas for elementary school teachers the shrinkage procedure sometimes increases the cross-year correlation but also sometimes decreases it. The difference comes about because middle school teacher effects have less variance after accounting for noise than do elementary school teachers. Consequently, the shrinkage is greater for middle school teachers and it reduces the influence teachers in the extremes of the distribution have on the correlation. The shrinkage tends to be smaller for elementary school teachers and has less impact on the correlations.

Following the previous literature on the stability of effects, we also track how the relative rankings of teachers change over time. Table 2 provides a tabulation of teacher rankings by quintile in the first two years, 2000/01 and 2001/02, for the largest school district in the sample, Hillsborough County. The table has raw counts of teachers in each cell, along with row percentages in brackets. The results are comparable to those reported by Aaronson, et al for Chicago and Koedel and Betts for San Diego. Table 3 provides one cell from the teacher ranking cross-tabulation, the proportion of teachers ranked in the top quintile in one year who are also ranked in the top quintile the following year. This is done for each county and each pair of years. In general, about one-third of teachers ranked in the top 20 percent one year are also ranked in the top quintile the following year. This proportion varies somewhat across districts and time periods, however, ranging from a low of 22 percent to a high of 47 percent.

B. Variance Decomposition

In order to gain a better understanding of the sources of variation in teacher effects, we empirically decompose the variance in the teacher effects derived from the baseline model after removing the variation due to independent student-level errors. We begin by distinguishing between the variance between and within teachers. We then break down the within-teacher variation into the components depicted in equation (5). Results are presented in Table 4.

The numbers in the first column of Table 4 indicate the absolute variance in teacher effects (expressed as average gains in test scores which are measured in standard deviation units) after accounting for independent student-level errors through maximum likelihood estimation.. We call this the total adjusted variance. For any given county, the numbers are two to six times larger in elementary school than in middle school, indicating much greater variation in measured teacher quality among teachers in the elementary grades.

Not only is the total adjusted variance in teacher effects greater in elementary school than in middle school, the proportion that is within teachers is much greater as well. With the exception of Orange County (where there is little difference between elementary and middle school), the proportion of variance that occurs within teachers is roughly twice as high at the elementary school level than at the middle school level.

The larger proportion of adjusted variance among estimated effects within elementary school teachers compared to middle school teachers does not appear to be due to greater variability in observed teacher, school, student, or peer effects. For Duval, Hillsborough and Palm Beach Counties we can explain about four to seven percent of the within-teacher variation in teacher effects in elementary schools with our observed variables. In contrast, for middle school teachers in those counties we can explain 15 to 37 percent of the within-teacher variance

in estimated teacher effects. A large portion of the adjusted variance remains between teachers and is unaccounted for amongst elementary school teachers. There are many possible sources for this variance such as true variability in teacher performance, omitted variables, aggregate shocks, or variance due to student teacher interactions that violate the assumption of additively separable effects.

C. Controls for Student, Peer and School Heterogeneity

In order to explore the effects of model specification on the stability of estimated teacher effects, we estimate a variety of achievement model specifications and then compute the year-to-year correlations in the resulting teacher effect estimates. Results are presented in Table 5. We begin by repeating the results of the baseline model that only includes student fixed effects in the first row. The second through fourth rows report two-year correlations of teacher effects derived from models in which observable time-varying student, peer and school characteristics are additively included in the achievement model. For both elementary and middle school, we see virtually no change in the inter-temporal stability of the teacher effect estimates as additional student, peer and school controls are added to the achievement model. This result is not surprising, given the very small portion of the adjusted within-teacher variance explained by these variables.

Rows 5 through 7 in Table 5 report cross-year correlations of teacher effects when either student heterogeneity is ignored entirely (row 5) or when student covariates, rather than student fixed effects, are used to control for student characteristics in the achievement model. Although the correlations remain virtually unchanged in the elementary school models for Duval, Hillsborough, and Orange counties, they increase somewhat in Palm Beach County for elementary school teachers. They increase substantially for middle school teachers in Duval

County, more than doubling when there is no adjustment for student heterogeneity. One likely explanation is that errors from selection are large when student fixed effects are omitted and the underlying rules for matching students to teachers are fairly stable over time. Thus the apparent increase in the inter-temporal stability of teacher effects we observe when student fixed effects are omitted is possibly a result of a stable bias in measured teacher effectiveness rather than stability in the true underlying teacher productivity. This suggests that achievement models that exploit student fixed effects to control for both observed and unobserved student characteristics that are time-invariant (Betts, et al. (2003), Clotfelter, et al (2007a, 2007b), Hanushek, et al., (2005)) may, in some cases, be less subject to bias than models that employ time-invariant student covariates to capture student heterogeneity (eg. Aaronson, et al. (2007), Kane, et al. (2006)).

In Hillsborough and Orange counties the correlations between middle school teacher estimates actually decrease substantially when student fixed effects are dropped from the model. This sort of precipitous drop in correlations is consistent with time-varying errors from selection in which there is heterogeneity in classroom assignments that is captured by the student fixed effects but this heterogeneity is not constant across years. For these conjectures to hold the assignment of students to classrooms must differ across counties and grade-levels. In Duval elementary schools, teachers must be consistently more likely to be assigned students who differ in terms of stable predictors of achievement. In Hillsborough and Orange counties there must be heterogeneity among classrooms on such factors but they are not consistently assigned across teachers. That is, there are relatively greater time-invariant errors from selection in Duval elementary classrooms and relatively greater time-vary errors from selection in Hillsborough and Orange middle school classroom. Regardless whether these conjectures are true or not, the

results suggest that in some cases controlling for fixed effects might improve the stability of teacher effects by reducing excess variability due to unmeasured student characteristics.

D. Variations in the Dependent Variable

The dependent variable employed in value-added analyses often varies across studies due to differences in scaling of test scores and differences in the test used to measure student achievement. Often researchers normalize test scores by grade and year in order to minimize the effects of differing exams or changes in the difficulty of exams over time. Although such normalizations are linear transformations of the underlying scale scores within a given grade-level, teacher effects combine data across grades so that inter-temporal correlations could be sensitive to such "norming." To gauge the impact of normalizing test scores we compare inter-temporal correlations of teacher effects from the baseline model with gains in normalized FCAT-NRT scores and an alternative model that uses gains in the raw FCAT-NRT scale scores in Table 6.[5]

Using the gains in the raw FCAT-NRT scores rather than the normalized scores has almost no appreciable effect on the inter-temporal correlation between 2001/02 and 2002/03. The difference in the correlations between estimates based on raw and normed scores typically 0.01 or 0.02 and 0.05 or less except for middle school teachers in Duval county were the correlation with the normed scores was 0.23, but the correlation with the raw scores was 0.31.

As noted by Lockwood et al. (2007) and Harris and Sass (2008a), estimates of teacher quality can vary depending on the exam used to measure student achievement. If different tests

[5] The FCAT-NRT is the Stanford achievement test, which is a vertically scaled exam that is designed so that a one-point change anywhere along the scale is equivalent. Consequently, it is valid to analyze changes in scale scores.

emphasize different kinds of material or have different effective maximums or "test ceilings" the measured effect of a teacher can vary depending on the test instrument being used. Likewise, if skills tested or ceiling effects change more often for one test relative to another, then inter-temporal stability of estimated teacher effects can vary across tests.

As indicated in Table 6, using gains in the normed FCAT-SSS, rather than gains in the normed FCAT-NRT, to estimate the teacher-by-year effects makes relatively little difference in the inter-temporal stability of the estimates for Hillsborough and Orange Counties. The cross-year correlations for 2001/02 and 2002/03 differed between the two sets of estimates by at most 0.08 for the two counties. In contrast, changing the test instrument produced some rather large differences in the inter-temporal stability of the teacher effect estimates for Duval and Palm Beach Counties. At the middle school level, the cross-year correlations in Duval and Palm Beach Counties were more than twice as high for the FCAT-SSS than for the FCAT-NRT. In contrast, at the elementary school level the cross-year correlation based on the FCAT-SSS was much higher in Palm Beach County but was substantially lower than that derived from the FCAT-NRT scores in Duval County. These results clearly suggest that using different tests can affect the stability of estimated teacher effects, but the cause of those differences is not clear. 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 in each year might provide insights into the difference we observe.

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 stability of

estimated teacher effects. In this paper we use a simple model for annual teacher effects and explored the inter-temporal stability of the resulting estimates based on different specifications of the model, different sample restrictions, and different test scores.

In general, we find that raw cross-year correlations are low to moderate, typically in the range of 0.2 to 0.3. The relatively low correlations appear to be driven in large measure by noise in teacher effect estimates. Dis-attenuated correlations, which correct for individual student-level errors in test performance, are much higher, typically in the range of 0.5-0.8. Overall the variability of teacher effects tends to be smaller for middle school mathematics teachers than for elementary grade teachers. The difference is not only a function of greater independent student-level errors in the elementary grade teacher estimates but also in the variance after accounting for such errors. In fact, in several instances the adjusted variance for middle school teachers is very small, about one sixth as large as the variance for elementary school teachers. Moreover, in three of the four counties included in this study, the year-to-year variability among estimates within teachers account for a much smaller share of the variance for middle school teacher than elementary school teachers.

The sources of the differences between elementary and middle school teachers are not clear. One possible explanation for the smaller overall variability in the effects of middle school teachers is that achievement growth is slower among older students. The marginal variance in achievement gains based on normed FCAT-NRT scores decreases with grade level and on average the variability in student gains in middle school grades is only about three quarters as large as it is for elementary school grades. However, even accounting for these differences, elementary school teacher effects are still 2.5 to 4.8 times as variable as their middle school

counterparts. Moreover a slowdown in growth does not explain why there is a greater share of the variance within elementary school teachers than for middle school teachers.

This difference across groups in the relative shares of different sources of variance also is not accounted for by the student, teacher, and school variables available for our analyses. These variables actually account for a smaller portion variance of the estimated effects for the elementary school teachers than for middle school teachers. It may be that classroom-level shocks are more common in elementary schools because teachers teach only one class or interactions between elementary school teachers and their students has greater effects on their annual performance. For example, elementary school students spend most of their day together with their classroom teacher which provides more time for interactions and an emergent class effect than in middle schools where students typically spend just one or a few periods with a teacher and change class groupings throughout the day so that an emergent classroom effect might have less opportunity to develop and even if they do develop they might get average out across the multiple classes middle school teachers teach. Alternatively, there may be more variation in unmeasured class-level supports, such as teacher aides or parent volunteers or it may even be related to differences between the elementary and middle school tests.

Regardless of the source, there are some important implications of the variance decomposition and the difference between elementary and middle school teachers. Given the large errors due to noise in student test scores, single-year estimates of teacher performance are unlikely to be sufficiently precise or stable across time to support performance pay or retention decisions. Given the decomposition results and the consistently low cross-year correlations from richer models, adjusting for observable variables does not appear likely to improve the estimates.

Averaging estimates across years is one potential means of reducing variation due to student-level independent errors. However, averaging estimates across years can introduce bias if true teacher performance varies across years. This makes averaging across years particularly appealing for reducing the variability from individual-level student errors in estimated effects for middle school teachers. For these teachers the independent student-level errors are relatively large and other sources of year-to-year variation are relatively small. Consequently, averaging is likely to yield substantial improvements to precision while introducing very little bias due to true variation in performance.

On the other hand, averaging estimated effects across years is somewhat less appealing for elementary school teachers. Averaging will still 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 is not guaranteed to be trivial as it appeared to be for middle school teachers. The mean squared error (MSE), 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.

Empirical Bayes shrinkage is another potential approach to reducing the variation due to independent student-level errors. EB shrinkage shrinks the estimates toward zero in proportion to a factor that depends on the ratio of variation in effects among teachers and the standard error of each individual teacher's estimated effect. EB shrinkage increases the precision of estimates, but again introduces bias into the estimated effects although MSE is still lower for the EB estimates than the raw estimates. As we discussed above, proper parameterization of the teacher effects to account for the various restrictions required for identification is essential for EB

estimation to produce interpretable results, e.g., estimate that do not shrink all teachers' estimate to an arbitrary holdout teacher. We applied EB shrinkage to our estimates and found that it had little effect on cross-year correlation for elementary school teachers but consistently resulted in increased correlations for middle school teachers. The differences were due in part to the fact that we estimated more variability among the true effects for elementary teachers than for middle school teachers resulting in more substantial shrinkage for middle school teachers.

EB shrinkage will limit the annual fluctuations of any given teacher's estimates and for teacher with small classes this reduction could be appreciable. However, as we found, it can have a limited effect on cross-year correlations. This occurs because the shrinkage factor applies to the noise or independent student-level errors and the true effects. Consequently, it can have less of an effect on the stability of a teacher's relative place in the distribution than on fluctuations in his or her yearly estimated effects. Averaging across years does not distort the stable teacher components and can improve the stability of a teachers relative position and the yearly estimates; however, it can come of the cost of smoothing away true annual fluctuations in performance.

An important finding from our study is the relatively limited impact of stratification on estimation and the inter-temporal stability of estimated effects. We are unaware of any previous research that explored the structure of the strata in student test score data and we consistently found that one large stratum containing 90 to 95 percent of teachers and students multiple small strata with only a handful of students and teachers (often just one) in each. These findings replicated across the counties used in this study as well as several other counties in Florida and other school districts. It is not clear how these results apply to data that includes students from multiple districts or greater geographic dispersion such a state. More work on the effects of

stratification remains. In particular, some teachers might be connected to the large stratum through very precarious links based on few students linked to few other students in the sample. Currently we know very little about how this might affect the precision of estimates or how to calibrate the strength of a teacher's connection to the large stratum.

However, even though stratification appears to be a relatively limited problem, careful parameterization of teacher effects is still useful for avoiding indeterminacy and instability of estimates due to arbitrary selection of holdout teachers. Our parameterization is not currently implemented in any readily available software including packages designed to model large-scale fixed effects problems such as the felsdvreg package for Stata (Cornelißen (2006)). We coded our models using SAS PROC GLM. This solution works well for small problems but becomes extremely memory intensive as the number of teachers gets large. For example, we were unable to apply our methods to data from Dade County because our 32 gigabytes of available RAM was insufficient. Iterative procedures could be adapted to use our parameterization but they do not currently provide standard errors. Similarly, the algorithm of Cornelißen (2006) could also be adapted to our parameterization and this is an important area for future research.

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.

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.

Betts, Julian R., Andrew C. Zau and Lorien A. Rice (2003). Determinants of Student Achievement: New Evidence from San Diego. San Diego: Public Policy Institute of California.

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.

Carlin, Bradley P., and Thomas A. Louis (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Clotfelter, Charles T., Helen F. Ladd and Jacob L. Vigdor (2007a). "How and Why Do Teacher Credentials Matter for Student Achievement?" Working Paper #2. Washington, DC: CALDER.

Clotfelter, Charles T., Helen F. Ladd and Jacob L. Vigdor (2007b). "Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects." Working Paper #11. Washington, DC: CALDER.

Cornelißen, Thomas (2006). "Using Stata for a Memory Saving Fixed Effects Estimation of the Three-Way Error Component Model." Unpublished. Hannover: University of Hannover.

Feng, Li and Tim R. Sass (2008). "Teacher Quality and Teacher Mobility." Unpublished. Tallahassee: Florida State University.

Hanushek, Eric A., John F. Kain, Daniel M. O'Brien, and Steven G. Rivkin (2005). "The Market for Teacher Quality." Working Paper #11154. Cambridge, MA: National Bureau of Economic Research.

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 (2008a). "The Effects of NBPTS Teachers on Student Achievement." Unpublished. Tallahassee: Florida State University.

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

Kane, Thomas J., Jonah E. Rockoff and Douglas O. Staiger (2006). "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City." Working Paper #12155. Cambridge, MA: National Bureau of Economic Research.

Kane, Thomas J., and Douglas O. Staiger (2008). Are Teacher-Level Value-Added Estimates Biased? An Experimental Validation of Non-Experimental Estimates. Unpublished. Cambridge, MA: Harvard University.

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., Daniel. F. McCaffrey, Laura S. Hamilton, Brian Stecher, Vi-Nhuan Le, , and Filipe Martinez (2007). "The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures," *Journal of Educational Measurement*, 44(1): 45-65.

McCaffrey Daniel F., Bing Han and J.R. Lockwood (2008). "From Data to Bonuses—A Case Study of the Issues Related to Awarding Teachers Pay on the Basis of Their Students' Progress." Working Paper #2008-14. Nashville, TN: National Center on Performance Initiatives.

Morris, Carl N. (1983). "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association* 78(381): 47-55.

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

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.

Table 1. Inter-temporal Correlations in Estimated Teacher-by-Year Average Effects (No Controls Other Than Student Fixed Effects)

		Varying Teachers Across 2-Year Periods			
		Correlation Between			
County	**Correlation Type**	**2000/01 and 2001/02**	**2001/02 and 2002/03**	**2002/03 and 2003/04**	**2003/04 and 2004/05**
		Elementary			
Duval	Raw	0.22	0.24	0.23	
	Dis-attenuated	0.45	0.49	0.48	
	Emp. Bayes	0.22	0.27	0.23	
Hillsborough	Raw	0.27	0.27	0.21	
	Dis-attenuated	0.59	0.55	0.44	
	Emp. Bayes	0.27	0.25	0.20	
Orange	Raw	0.29	0.34	0.17	
	Dis-attenuated	0.79	0.74	0.77	
	Emp. Bayes	0.32	0.37	0.28	
Palm Beach	Raw	0.21	0.09	0.22	
	Dis-attenuated	0.33	0.17	0.81	
	Emp. Bayes	0.17	0.10	0.28	
		Middle			
Duval	Raw	0.22	0.09	0.05	0.18
	Dis-attenuated	0.67	0.76	0.76	0.62
	Emp. Bayes	0.28	0.37	0.27	0.25
Hillsborough	Raw	0.19	0.26	0.27	0.29
	Disattenuated	0.80	0.76	0.79	1.00
	Emp. Bayes	0.37	0.31	0.34	0.24
Orange	Raw	0.35	0.27	0.26	0.11
	Disattenuated	0.91	0.92	1.00	0.64
	Emp. Bayes	0.27	0.27	0.29	0.18
Palm Beach	Raw	0.30	0.19	0.14	0.08
	Disattenuated	0.99	1.00	0.80	0.52
	Emp. Bayes	0.26	0.32	0.35	0.16

Table 2. Quintile Ranking of Estimated Teacher-by-Year Average Effect in 2001/02 by Quintile Ranking in 2000/01, Hillsborough County (No Controls Other Than Student Fixed Effects)

Elementary (Cross-Year Correlation = 0.27)

Quintile in 2000/01	Quintile in 2001/02					
	1	**2**	**3**	**4**	**5**	**Total**
1	34 [27.9]	27 [22.1]	21 [17.2]	23 [18.9]	17 [14.0]	122 [100.0]
2	26 [22.4]	32 [27.6]	26 [22.4]	19 [16.4]	13 [11.2]	116 [100.0]
3	29 [22.0]	30 [22.7]	23 [17.4]	26 [19.7]	24 [18.2]	132 [100.0]
4	25 [17.6]	26 [18.3]	29 [20.4]	24 [16.9]	38 [26.8]	142 [100.0]
5	11 [8.0]	21 [15.3]	31 [22.6]	35 [25.6]	39 [28.5]	137 [100.0]
Total	125 [19.3]	136 [21.0]	130 [20.0]	127 [19.6]	131 [20.2]	649 [100.0]

Middle (Cross-Year Correlation = 0.19)

Quintile in 2000/01	Quintile in 2001/02					
	1	**2**	**3**	**4**	**5**	**Total**
1	13 [21.3]	16 [26.2]	16 [26.2]	7 [11.5]	9 [14.8]	61 [100.0]
2	10 [16.7]	13 [21.7]	16 [26.7]	15 [25.0]	6 [10.0]	60 [100.0]
3	7 [11.3]	11 [17.7]	16 [25.8]	18 [29.0]	10 [16.1]	62 [100.0]
4	4 [6.8]	8 [13.6]	13 [22.0]	20 [33.9]	14 [23.7]	59 [100.0]
5	7 [10.9]	6 [9.4]	9 [14.1]	21 [32.8]	21 [32.8]	64 [100.0]
Total	41 [13.4]	54 [17.7]	70 [22.9]	81 [26.5]	60 [19.6]	306 [100.0]

Note: row percentages in brackets.

Table 3. Percentage of Teachers Who Remain in Top Quintile from One Year to the Next Based on Teacher-by-Year Average Effects (No Controls Other Than Student Fixed Effects)

Elementary

County	2000/01 and 2001/02	2001/02 and 2002/03	2002/03 and 2003/04	2003/04 and 2004/05
Duval	26.6	30.9	34.8	
Hillsborough	28.5	25.7	24.5	
Orange	29.2	35.5	32.8	
Palm Beach	33.9	22.0	34.1	

Middle

County	2000/01 and 2001/02	2001/02 and 2002/03	2002/03 and 2003/04	2003/04 and 2004/05
Duval	37.0	25.0	34.8	28.0
Hillsborough	32.8	28.4	30.6	34.1
Orange	22.9	29.0	36.1	26.3
Palm Beach	47.4	31.3	29.6	31.4

Table 4. Variance Decomposition of Estimated Teacher-by-Year Average Effects

County	Total Variance in Teacher-by-Year Effect	Percent of Variance Between Teachers	Percent of Variance Within Teachers	Percent of Within-Teacher Variance Due to Variation in Student/ Peer Charact-eristics	Percent of Within-Teacher Variance Due to Variation in Teacher Charact-eristics	Percent of Within-Teacher Variance Due to Variation in School Charact-eristics
			Elementary			
Duval	0.0610	30.0	70.0	1.7	1.5	0.4
Hillsborough	0.0358	48.0	52.0	2.8	3.4	0.5
Orange	0.0459	76.5	23.5	17.5	11.2	6.1
Palm Beach	0.0482	39.4	60.6	2.0	3.0	0.1
			Middle			
Duval	0.0127	65.9	34.1	17.6	5.8	1.6
Hillsborough	0.0116	69.8	30.2	5.1	8.5	0.9
Orange	0.0075	77.8	22.2	2.8	16.5	24.9
Palm Beach	0.0088	64.5	35.5	17.0	13.6	6.0

Table 5. Inter-temporal Correlation of Estimated Teacher-by-Year Average Effects in 2000/01 and 2001/02 with Alternative Student, Peer and School Controls (No Minimum Students-Per-Teacher-Year Restriction)

					Counties			
Outcome	Student Time-Invariant Controls	Student Time-Varying Controls	Peer Time-Varying Controls	School Time-Varying Controls	Duval	Hills-borough	Orange	Palm Beach
Elementary								
Gain on Normed FCAT-NRT	Student Fixed Effects	No	No	No	0.22	0.27	0.29	0.21
Gain on Normed FCAT-NRT	Student Fixed Effects	Yes	No	No	0.22	0.27	0.30	0.21
Gain on Normed FCAT-NRT	Student Fixed Effects	Yes	Yes	No	0.22	0.25	0.27	0.19
Gain on Normed FCAT-NRT	Student Fixed Effects	Yes	Yes	Yes	0.23	0.27	0.29	0.19
Gain on Normed FCAT-NRT	None	No	No	No	0.22	0.26	0.26	0.28
Gain on Normed FCAT-NRT	Student Covariates	No	No	No	0.22	0.26	0.28	0.27
Gain on Normed FCAT-NRT	Student Covariates	Yes	Yes	Yes	0.23	0.27	0.28	0.29
Middle								
Gain on Normed FCAT-NRT	Student Fixed Effects	No	No	No	0.22	0.19	0.35	0.30
Gain on Normed FCAT-NRT	Student Fixed Effects	Yes	No	No	0.22	0.19	0.35	0.31
Gain on Normed FCAT-NRT	Student Fixed Effects	Yes	Yes	No	0.24	0.17	0.36	0.32
Gain on Normed FCAT-NRT	Student Fixed Effects	Yes	Yes	Yes	0.25	0.18	0.35	0.32
Gain on Normed FCAT-NRT	None	No	No	No	0.44	0.07	0.10	0.28
Gain on Normed FCAT-NRT	Student Covariates	No	No	No	0.42	0.07	0.10	0.26
Gain on Normed FCAT-NRT	Student Covariates	Yes	Yes	Yes	0.34	0.08	0.09	0.25

Table 6. Inter-temporal Correlation of Estimated Teacher-by-Year Effects in 2001/02 and 2002/03 Using Alternative Test Scores (No Minimum Students-Per-Teacher-Year Restriction)

					Counties			
Outcome	Student Time-Invariant Controls	Student Time-Varying Controls	Peer Time-Varying Controls	School Time-Varying Controls	Duval	Hills-borough	Orange	Palm Beach
Elementary								
Gain on Normed FCAT-NRT	Student Fixed Effects	Yes	Yes	Yes	0.26	0.29	0.28	0.07
Gain on Normed FCAT-SSS	Student Fixed Effects	Yes	Yes	Yes	0.11	0.35	0.23	0.42
Gain on FCAT-NRT Scale Score	Student Fixed Effects	Yes	Yes	Yes	0.21	0.29	0.29	0.09
Middle								
Gain on Normed FCAT-NRT	Student Fixed Effects	Yes	Yes	Yes	0.23	0.22	0.38	0.30
Gain on Normed FCAT-SSS	Student Fixed Effects	Yes	Yes	Yes	0.56	0.27	0.26	0.61
Gain on FCAT-NRT Scale Score	Student Fixed Effects	Yes	Yes	Yes	0.31	0.24	0.33	0.30

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