The availability of administrative data on teachers and students has greatly enhanced the ability of researchers to address research topics related to the effectiveness of teachers. Such data permit the researcher to use the student as the unit of observation, to follow students over time, and in many cases to match students with their specific teachers. Moreover, the sample sizes are sufficiently large to allow for more sophisticated and complex modeling than has heretofore been possible. Now that No Child Left Behind (NCLB) legislation requires every state to test all students in Grades 3-8 annually, the hope is that administrative data will become even more readily available for research on teachers.

Among the many issues that arise in estimating the effectiveness of teachers, three are particularly salient. One reflects the observation that teachers are not randomly assigned to schools or to classrooms within schools. As a result, teacher effects may be confounded by the unmeasurable characteristics of students, such as their ability and motivation. Any estimates of teacher effects that do not fully account for the nonrandom matching of students to teachers would be biased upward if students with greater ability and motivation are assigned to the more effective teachers, and the effects would be downward biased if school or district administrators try to compensate for lower student ability by assigning them to the more effective teachers. This nonrandom assignment of teachers and students represents a significant obstacle to the estimation of teacher effects.

A second issue relates to the technical and conceptual feasibility of separating the effects of individual teachers (or teacher credentials) from the effects of other inputs to the educational process, such as the characteristics of students in the classroom, school level policies, and characteristics of the individual students. A third concern arises because measurement error leads to imprecise estimates of teacher effects and complicates their interpretation.

Most researchers now agree that teacher quality matters for student achievement and that variation in teacher effectiveness contributes significantly to the variation in student achievement. Part 1 briefly reviews the evidence for that conclusion. That conviction, combined with the current national focus on student achievement as exemplified by the federal No Child Left Behind Act of 2001, has encouraged
policymakers in some states and districts to introduce programs to reward individual teachers for their effectiveness in raising student test scores.

As discussed in Part 2, however, this policy thrust appears to be moving faster than can be supported by the technical analysis of teacher effects. Much remains to be learned about both the estimation of teacher effects and their usefulness for policy.

Part 3 shifts the focus to teacher credentials. Although many researchers and policymakers have traditionally downplayed the relationship between teacher credentials and student achievement, some researchers, including me, believe that teacher credentials are important predictors of student achievement and should be viewed as important policy levers for improving student achievement and for reducing achievement gaps.

---

**Part 1**

**Do Teachers Matter?**

Only recently have researchers documented in a reasonably convincing way what parents always knew, namely that the variation in teacher quality contributes significantly to the variation in student outcomes. Prior to this recent research, the more standard view among many researchers, which was based on early studies showing little relation between teacher credentials and student achievement, was that variation across teachers does not account for much of the variation in student achievement or achievement gains. The conclusion that teachers matter is based on three quite different approaches to isolating the effects of teachers.

One approach emerges in work by Hanushek and Rivkin and various coauthors using Texas data. In Rivkin, Hanushek and Kain (2005), the authors use statewide data on student test scores that can be matched to teachers only at the grade level. Although their approach is clever, more convincing evidence requires that the teachers of students be identified at the classroom level. In Hanushek, Kain, O’Brien, and Rivkin (2005) the authors use data for a single Texas district, which they refer to as the Lone Star District, for which they are able to match students in Grades 3–8 to their classroom teachers. In both cases, the authors’ goal is to measure the persistent component of teacher effects, that is, the nonrandom component of what is undoubtedly a noisy measure.

The patterns that emerge from the latter study are summarized in Table 1. In all cases the estimates are based on teacher effects by year, with the achievement gains measured in normalized units. The table highlights three important issues that arise in the estimation of such effects. One issue is the extent to which controls are included for the demographic characteristics of the students such as their income, race/ethnicity and gender. Another is whether the estimates refer to the variation in teacher quality at the district level or to the variation within individual schools. The third issue is the importance of measurement error.

The first row of Table 1 summarizes the variation in student achievement gains accounted for by the variation in teacher effects (estimated using fixed effects for teachers by year, a method that is discussed further below) for each of four models:
district-wide models with and without demographic adjustments and models that include school-by-year fixed effects with and without demographic adjustments. Not surprisingly, the within-school variation in teacher effects is smaller than the variation across the district. Further, as would be expected, demographic controls reduce the estimated variation in teachers effects far more for the within-district estimates than for the within-school estimates. That outcome reflects the fact that more student and teacher sorting occurs across schools than within them.

Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Within district</th>
<th>Within school and year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unadjusted</td>
<td>Demographic controls</td>
</tr>
<tr>
<td>Teacher-Year Variation(^{a})</td>
<td>0.210</td>
<td>0.178</td>
</tr>
<tr>
<td>Adjacent Year Correlation</td>
<td>0.500</td>
<td>0.419</td>
</tr>
<tr>
<td>Teacher Quality Variance</td>
<td>0.105</td>
<td>0.075</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>(0.32)</td>
<td>(0.27)</td>
</tr>
</tbody>
</table>

Source: Hanushek, Kain, O’Brien, and Rivkin (2005), Table 1.

\(^{a}\) The entries in this row are the variance in student achievement gains explained by fixed effects for teachers by year.

\(^{b}\) The demographic controls include free or reduced lunch, gender, race/ethnicity, grade, limited English proficiency, special education, student mobility status, and year indicator variables.

Those entries overstate the contribution of teachers to student achievement, however, because some of the variation is simply measurement error in the form of random noise. The entries in the second row of Table 1, which are the year-to-year correlation in teacher effects, indicate that from 42 to 50 percent of the variation is persistent. These fractions are used to adjust the entries in the first row downward to generate the variance in persistent variation in teacher quality reported in the third row for each of the four models. Below each measure of variance is a standard deviation, calculated as the square root of the corresponding variance. These estimates imply that a one standard deviation in teacher quality is associated with a 0.22 to 0.32 standard deviation difference in achievement gains. The larger estimate could well overstate the importance of teachers since it does not control for school level factors such as the effectiveness of school principals or the composition of the students. Hence, the authors highlight the smaller estimate, emphasizing it is a lower bound estimate of teacher effects.\(^{1}\)
A second approach based on national data is presented in Rowan, Correnti, and Miller (2002). These authors study two cohorts of students in the nationally representative sample of schools in the Prospects study. The authors fit four different models for each subject, cohort and grade. The first model is a three-level, nested analysis of variance with students clustered within classrooms (or teachers), classrooms within schools, and schools. The second and third are value-added and gains models similar to those described below. The fourth is a cross-classified model. In none of the models are the authors able to separate teacher effects from classroom effects.

For each of their models, the authors find that classrooms (and their teachers) account for significant portions of the relevant variance in achievement, where the relevant variance is defined in different ways in the various models. For example, on the basis of their fourth model, the authors conclude that the variability in teacher effects accounts for 60 to 61 percent of the “reliable” variance in achievement growth in reading, and 52 to 72 percent in math, where the “reliable” variance is defined as the variability in achievement growth purged of the effects of measurement error. In a subsequent review of this study, McCaffrey Lockwood, Koretz, and Hamilton (2003) quibble with the way the authors measure reliable variance. Nonetheless, they conclude that this study provides convincing evidence of teacher effects, or more precisely, classroom effects, but that the magnitudes are not fully clear.

Additional evidence emerges from a study that uses data from the Tennessee class-size experiment examining teacher effects for students in the early grades (Nye, Konstantopoulos, and Hedges, 2004). This mid-1980s experiment provides the only evidence about teacher effects based on a random assignment of students to teachers. Teacher effects are estimated using a hierarchical linear model designed to sort out the between-teacher (but within-school) effects of teachers on achievement gains and also on achievement status. The authors conclude that the teacher effects are real and consistent with those of other studies that they are larger for math than for reading, and that within-school teacher effects are larger than across-school effects. For math, the estimated between-teacher variance components range from 0.123 to 0.135, and for reading the effects are about half that size. If teacher effects are normally distributed, these findings suggest that the difference between having a teacher at the twenty-fifth percentile and the seventy-fifth percentile is nearly one-half a standard deviation in math and nearly one-third a standard deviation in reading.

Taken together, these and other studies provide convincing evidence that teachers matter for student achievement. Not examined here are other studies showing that these teacher effects accumulate over time (McCaffrey et al., 2003, pp. 36-48). Although the overall contribution of teachers to student achievement has not been precisely established, the findings in these papers are sufficient to justify additional research attention to other questions related to teachers, such as whether it is possible to identify the effectiveness of individual teachers and whether teacher credentials are predictive of student achievement.
Can Teacher-Specific Effects Be Identified and Measured?

Identifying the relative effectiveness of individual teachers is of increasing policy relevance as policymakers explore the idea of rewarding individual teachers for good performance, as measured by their ability to raise test scores. The research in this section shows, however, that it is difficult to separate the effects of teachers from other inputs, particularly those deriving from contextual factors within the school or in the classroom. In addition, it shows that the estimated teachers effects are not very stable over time, particularly if they are not adjusted for measurement error, and that there is no clear best way to deal with measurement error in the estimates.

Researchers have been using two main approaches to identify the effectiveness of individual teachers in raising student achievement. I refer to the first approach as “value-added modeling” and include in that category both levels and gains models. The second approach includes mixed and layered models that directly model the full joint distribution of all student outcomes. Although for some purposes the mixed-methods models are superior, they are computationally very demanding and receive somewhat less attention in this overview.

VALUE-ADDED MODELS

As noted earlier, a fundamental challenge in estimating teacher effects is the observation that students are not randomly assigned to teachers. For the moment, I set this issue aside to develop the conceptual foundations of the standard value added model, with particular attention to the assumptions underlying it.

Derivation of Simple Value-Added Model

The starting point for this analysis is the observation that education is a cumulative process. In the context of a very simple model in which the only educational input that matters is teachers and in which all other determinants of achievement such as student background, ability, and motivation are set aside, we can write:

\[ A_{it} = \phi(T_{it}, T_{i,t-1}, \ldots) + \epsilon_{it} \]  

[equation 1]

where:

- \( A_{it} \) refers to student \( i \)'s achievement, as measured by test scores, in year \( t \),
- \( T_{it} \) refers to some measure of the quality of student \( i \)'s teacher in year \( t \), and
- \( \epsilon_{it} \) is an error term.

This equation expresses student \( i \)'s achievement in year \( t \) as a function of her teacher in that year and in all previous school years plus a random error.

Two additional assumptions permit this relationship to be transformed into one that can be used to estimate the effect of the student’s teacher in year \( t \) on the
student’s achievement in that same year, controlling for the effects of teacher quality in all prior years. One assumption is that the relationship is linear and that the teacher quality measure has a constant marginal impact on student achievement. The second is that student achievement, or knowledge, decays from one year to the next at a constant rate. As a result, the rate at which a student’s knowledge persists from one year to the next is also constant. Letting $\beta$ be the effect of $T$ and $\alpha$ the rate at which knowledge persists, we can rewrite equation 1 as:

$$A_{it} = \beta T_{it} + \alpha \beta T_{it-1} + \alpha^2 \beta T_{it-2} + \ldots + \varepsilon_{it} \quad [equation 2]$$

and, after rearranging terms, as:

$$A_{it} = \beta T_{it} + \alpha (\beta T_{it-1} + \alpha \beta T_{it-2} + \alpha^2 \beta T_{it-3} + \ldots ) + \varepsilon_{it} \quad [equation 3]$$

Noting that the expression within the parentheses is simply $A_{it-1}$ and changing the order of the terms, we end up with:

$$A_{it} = \alpha A_{it-1} + \beta T_{it} + u_{it}, \quad [equation 4]$$

where the error term, $u_{it}$, is equal to $\varepsilon_{it} - \alpha \varepsilon_{it-1}$.

Thus, the effects on current achievement of the student’s prior teachers are captured by the lagged achievement term. If a student’s knowledge does not persist from year to year, the persistence parameter, $\alpha$, would be zero.

Models of this form (but with additional explanatory variables as discussed below) are typically referred to as “value-added models” and are commonly used to estimate $\beta$, namely, the effect of current teachers on current achievement. The popularity of such models derives largely from their simplicity and intuitive appeal. Logically, it makes sense to control statistically for the achievement, or knowledge, that the student brings to the classroom at the beginning of the year when estimating the effect of his or her current teacher. In addition, the value-added model is flexible in that it does not impose a specific assumption about the rate at which knowledge persists over time; instead, it allows that rate to be estimated. Nonetheless, the model is valid only if the underlying assumptions about the constancy of effects are valid. Further such models raise statistical concerns because of the inclusion on the right hand side of the equation of the lagged achievement term, which in the presence of serial correlation would be correlated with the error term.

**Gains Model**

This last statistical problem can be avoided by assuming there is no decay in knowledge so that the persistent parameter, $\alpha$, equals 1 and moving the lagged achievement term to the left hand side of the equation. This procedure generates the gains model:

$$A_{it} - A_{it-1} = \beta T_{it} + \varepsilon_{it}. \quad [equation 5]$$

In this case, the parameter, $\beta$, refers to the effect of a student’s teacher on his or her gains in achievement. If the assumptions underlying the initial value-added model are correct, however, and the decay rate is not zero, the gains model is incorrectly specified. The reason is that the term $(\alpha - 1)A_{it-1}$ is now missing from the right hand side of the equation. To the extent that prior year achievement is positively correlated with teacher effects, the estimated teacher effects would be biased
downward. Thus, within the framework of education as a cumulative process, the shift to the gains model solves one statistical problem but introduces a new one.

**Full Value Added (or Gains) Model with Student Fixed Effects**

In fact, most researchers estimate a richer form of the simple model in equation 1, one that includes time-varying student characteristics, classroom or school characteristics, and student fixed effects. This full model can only be estimated with longitudinal data on individual students and multiple cohorts of students. If data are available for only a single cohort of students, no classroom characteristics such as class size or the composition of the students, can be included in the equation because teachers and their classrooms are indistinguishable.

\[ A_{it} = \alpha A_{i,t-1} + \beta T_{it} + \gamma X_{it} + \delta C_{it} + \theta_i + \eta_{it} \]  

*equation 6*

where:

- \( A_{it}, A_{i,t-1} \) and \( T_{it} \) are as defined above, and
- \( X_{it} \) are time varying student variables
- \( C_{it} \) are classroom and school characteristics in year \( t \)
- \( \theta_i \) are student fixed effects
- \( \eta_{it} \) is an error term.

For this model to be consistent with the cumulative model of the education process, the same assumptions that were needed to derive the simple value added model in equation 4 are needed. In particular, each of the variables must exert a constant linear effect on student achievement in each year and their effects on student achievement must all decay at the same rate \((1-\alpha)\).

The student fixed effects are a crucial part of this enriched model. They control for the time-invariant characteristics of students—both those that are measurable and those that are not—and under certain assumptions address the fundamental problem highlighted above, namely that the teachers are not randomly assigned to students. The inclusion of student fixed effects means that the teacher effects are derived from the within-student variation in student achievement. The key assumption needed for student fixed effects to address fully the concern about nonrandom sorting is that students are assigned to teachers on the basis of their permanent or average characteristics rather than on any time-varying, unmeasurable characteristics. Most value-added studies of teacher effects either implicitly or explicitly make this assumption. I return below to Jesse Rothstein’s recent test of the validity of this assumption.

In the context of these models, the teacher variables are typically entered as 0-1 indicator variables, either for each teacher or for each teacher by year. Thus, teacher effects are estimated by the method of teacher fixed effects (in contrast to the method of random effects), an approach that seems reasonable given the goal of determining the effectiveness of a group of specific teachers.

Two issues arise in estimating and interpreting such models. One is the technical challenge of using a program such as *Stata* to generate teacher effects in a model
that also includes student fixed effects. Although Stata can easily handle one set of fixed effects through the process of demeaning (for example, by subtracting the mean value for each student from all the variables in the model), it cannot use that procedure simultaneously for a second set of fixed effects. A natural solution to that technical problem is to create a new set of indicator variables that combine the student and teacher indicator variables into a single set of student-teacher indicator variables. Although that process works well for some purposes, it makes it difficult to capture the individual teacher effects. New programs are becoming available to address this technical problem (Cornelißen 2006). An alternative solution to this problem is to replace the student fixed effects with a vector of student characteristics. The shortcoming of that approach is that it ignores the unmeasurable characteristics of students, some of which could well be correlated with teacher quality.

The second and more important issue relates to measurement error. Note that teacher-by-year fixed effects are identified by the number of students taught by the teacher in that year, a number that could well be quite small, especially at the elementary school level. Even when teacher effects are based on multiple years of data, the number of students taught will differ across teachers, which means that the coefficients of the teacher indicator variables are estimated with different degrees of precision. Had they been estimated by random effects rather than by fixed effects, estimates for individual teachers would have been shrunken toward the mean, with the amount of shrinkage greater for the teacher effects that are estimated with less precision.

Letting $\beta_t^*$ represent the predicted teacher effect for teacher $t$ that emerges from a fixed effects specification, $\beta$, the true value and $\epsilon$ a random error, we can express the predicted teacher effect that emerges from a fixed effect specification as a function of the true effect plus an error term as follows:

$$\beta_t^* = \beta_t + \epsilon$$  \hspace{1cm} [equation 7]

One might then calculate the true teacher effect for any given teacher as a weighted average of the estimated teacher effect for that teacher and the mean teacher effect for the sample as a whole:

$$\lambda \beta_t^* + (1-\lambda) \text{mean } B_t^*$$  \hspace{1cm} [equation 8]

where:

$$\lambda = \frac{\text{Var}\beta_t}{(\text{Var}\beta_t + \text{Var}\epsilon)}.$$  

Thus, the larger is the random error of the estimate, the smaller is $\lambda$ and the greater the weight placed on the mean teacher effect. Although such an adjustment is conceptually straightforward, estimating $\lambda$ directly from the variances can be tricky to implement in practice. One implication of this shrinkage procedure is that teachers who teach small numbers of students are unlikely to be identified as either particularly effective or particularly ineffective teachers. Although the outcome on the low side may be appropriate given that it would protect decent teachers with small classes from being unjustly sanctioned, the shrinkage procedure could also keep some very effective teachers from being recognized.
Additional Considerations

Although much more could be said about this standard value added (or gains model), I add here only two additional considerations. The first refers to the role of parents. As pointed out by Todd and Wolpin (2003), compensating behavior by parents could potentially mute the estimated differences in teacher effectiveness. That outcome would occur if parents spend more productive time working on school work with their children when their children have ineffective teachers than when they have effective teachers.

Another consideration is whether to include school fixed effects in the model. Often they are not included, particularly if student fixed effects are in the model, as in equation 6. In the absence of student fixed effects, the addition of school fixed effects can help mitigate the problem caused by the non-random assignment of teachers to students. Their inclusion in the model means that teacher effects are identified solely by differences in teacher quality within schools. As a result, the estimates of teacher effects are not contaminated by the fact that the more effective teachers are more likely to end up in schools with the more able and more motivated students. Including school, rather than student fixed effects, however, does not account for the possibility that the more able students within a school may be assigned to the higher quality teachers. At the same time, their inclusion means that a teacher’s effectiveness is measured relative to other teachers in the school rather than to a broader set of teachers. As emerged from Table 1 above, the overall estimated variation in teacher effectiveness will be smaller when fixed effects are included for schools than when they are excluded.

How Stable Are Teacher Effects?

In most cases one would expect that a teacher who is very effective (or ineffective) in one year would be similarly effective (or ineffective) in the following year. Hence, one way to evaluate the validity of the teacher effects that emerge from value added models would be to examine their stability from one year to the next. The more unstable they are, the less useful they are likely to be in making high-stakes decisions about teachers.

Only a few studies have explored the stability of teacher effects (Ballou, 2005; Aaronson, Barrow, and Sander, 2007; Koedel and Betts, 2007). In all cases, the studies find that teacher effects are quite unstable. For example, Koedel and Betts (2007) rank teachers in San Diego by their estimated fixed effects for two years in a row. They find that among those who are ranked in the lowest quintile in the first year, only 30 percent stay in that quintile in the next year and another 31 percent move up to one of the top two quintiles. A similar pattern emerges at the top of the distribution. Although 35 percent of teachers who are initially ranked in the top quintile remain there in the second year, 30 percent fall to the first or second quintile (cited in Lockwood, McCaffrey, and Sass, 2008).

The most complete study of the stability of teacher effects is by Lockwood et al. (2008). The authors focus on math teachers because teacher effects are generally larger for math than for reading. They start with a very simple gains model—one with student fixed effects and teacher-by-year fixed effects—and then examine how modifying the model changes the results. They estimate all models at the district level and do not include school fixed effects. Thus, the teacher effects are
measured relative to the average of all teachers in the district in the relevant subject and grade range, not relative to the average teacher at a given school.

The correlations of teacher fixed effects for middle-school teachers across adjacent years in each district are moderately low, typically in the range of 0.2 to 0.4, and do not change much as the model is enriched with additional covariates or modified. The correlations of teacher effects across adjacent years for elementary school teachers are even lower, typically in the range of 0.1 to 0.3. Much of the apparent instability is attributable to the noise in student test scores. After the authors adjust the estimates for measurement error, the correlations increase significantly, up to the 0.5 to 0.8 range for both elementary and middle-school teachers (Lockwood et al., 2008, Table 3).

The authors tried to determine the causes of the instability by examining how the results were affected by the use of normalized or non-normalized test scores, the extent to which teachers have some students in common, and the addition of covariates to the value-added model. With a few minor exceptions, the instability of the effectiveness rankings was not very sensitive to the various changes. The authors end on a cautionary note. Although the measures adjusted for measurement error could well be acceptable for some decisions, the raw, or unadjusted, measures are too unstable to be used for high-stakes personnel decisions.

The Rothstein Challenge

Another challenge to the validity of the value-added approach to estimating teacher effects appears in a recent paper by Rothstein (2007). As emphasized above, one of the advantages of longitudinal data sets for estimating teacher effects is that they permit the researcher to use student fixed effects to control for the time-invariant, student-level characteristics, both measured and unmeasured, that may be correlated with the teacher measures. The inclusion of fixed effects for students solves the problem of the nonrandom matching of students to teacher when such matching is based on the time-invariant characteristics of the students, such as their basic ability or motivation. Rothstein refers to such matching as “static tracking” and contrasts it to the “dynamic tracking” that occurs when school administrators sort students into classrooms and teachers in a nonrandom way on the basis, in part, of students’ current performance.

He correctly emphasizes the importance of testing the assumption of static tracking and does so by introducing a placebo. In particular, using data for one cohort of elementary school students in North Carolina, he estimates a value added model that includes not only the student’s current teacher (e.g., the student’s fourth-grade teacher) but also his or her subsequent teacher in the following grade (fifth grade). If the basic value-added model is correct, the fifth-grade teacher should have no impact on the student’s fourth-grade test scores (or more precisely in the context of a model with student fixed effects, on the extent to which the student’s fourth-grade test score deviates from his or her average test score). In fact, however, Rothstein finds that the student’s fifth-grade teacher has almost as significant an impact on his or her fourth-grade scores (in reading) as does the fourth-grade teacher. This occurs, he argues, because the student’s fourth-grade test score is used to determine his or her fifth-grade teacher.
If Rothstein is correct about the importance of dynamic tracking, his analysis represents a serious challenge to the validity of the standard value approach. The argument on the face of it seems quite compelling. At the same time, it appears to imply that all the estimated teacher effects are spurious, which conflicts with the conclusion from other studies that teachers matter. Hence additional research on the validity of the static tracking model is clearly needed.

A first step would be to re-estimate the Rothstein models with multiple cohorts, and to examine results for math in addition to reading. The use of multiple cohorts would permit the researcher to separate teacher effects from contextual effects, which as discussed below, have emerged as a cause of concern with respect to the estimation of teacher effects in more complex models. Although Rothstein believes that the use of multiple cohorts will not change the results (personal communication with the author, April 2008), it would be useful to have that confirmed empirically.

A second step would be to explore the student-assignment process used by school principals. My preliminary investigation of this matter in a few North Carolina elementary schools offers little support for the hypothesis of dynamic tracking in some schools, but my investigation was limited. Clearly, more investigation is needed. In addition, it might be productive to remove the school fixed effects from the Rothstein’s equation to estimate teacher effects relative to teachers throughout the district rather than to those within each school.6

Mixed Methods or Layered Models (Multivariate Modeling)

These models are far more complicated than the simple value added models in that they specify a joint distribution for the entire multivariate vector of test scores.7 Included among these models are the Tennessee Value Added Assessment System (TVAAS) developed by William Sanders for Tennessee; the cross-classified models of Rowan, Correnti, and Miller (2002) and Raudenbush and Bryk (2002); and the persistence models of McCaffrey et al. (2003).

The key element of such models is that a student’s performance in any year is modeled not only as a function of his or her teacher in the current year, but also of teachers in prior years. Moreover, in such models, teacher effects are typically estimated using random, rather than fixed, effects. A major advantage of multivariate models relative to the simpler value added models with fixed effects is that the models use more information to identify teacher effects. In particular, they incorporate student scores in later years, which hold information about the effectiveness of teachers in the past. Another advantage is that they are very flexible. The primary disadvantage of such models is their tremendous computational demands. Until computational methods are developed to make it easier to estimate such models, it is likely that the more standard value-added models will be the basis of much of the ongoing research in this field.

I focus here on the TVAAS layered model because it has received significant attention in the literature. Implicit in this specific model is the assumption that any teacher effects in a prior year persist undiminished in future years. No student covariates are included. Instead, the complex correlations among the errors from the repeated test scores substitute for student specific covariates.
Kupermintz (2003) provides some useful insights into the TVAAS model. First the resulting estimates rank teachers within each school system. Hence, a weak teacher in a system with many other weak teachers may receive a more favorable ranking that a similar teacher in a stronger system. Second, the teacher effects are “shrunk” towards the system average for reasons similar to those discussed above. Thus, once again, it is difficult to get accurate estimates of the effectiveness of teachers who are working with small numbers of students.

In addition, and perhaps most significantly, Kupermintz questions the validity of the estimated teacher effects given that they emerge from a model that includes no student level or classroom level covariates. Although he acknowledges that the model uses prior achievement as a covariate or “blocking variable,” which means that each child serves as his or her own control, he notes that such “blocking” procedures were developed in the context of controlled experiments not in the context of observational studies. In contrast to controlled experiments in which treatments can be randomly assigned, students are not randomly assigned to teachers (Kupermintz, 2003, p. 292). As a result, the estimated teacher effects may be confounded by the effects of correlated student level characteristics that are omitted from the model. Further, he argues that for the TVAAS procedure to be valid, the prior year achievement variable would have to serve as a proxy for a variety of contextual factors including, for example, the socio-economic or achievement mix of students in the classroom.

Lockwood and McCaffrey (2007) have examined the extent to which the absence of covariates, at either the student level or at the classroom level, distorts the results. They examined this in the context of a general multivariate model (also see McCaffrey, Lockwood, Koretz, Louis, and Hamilton, 2004). Despite concerns that the use of random effects models can lead to inconsistent estimates when unobserved individual effects are correlated with other variables in the model, Lockwood and McCaffrey (2007) demonstrate through analysis and simulation that the mixed-method approach does not generate much bias in practical applications, particularly when the number of tests scores for individual students is relatively large.

The authors’ simulations support the claim of Sanders that the joint estimation of multiple test scores for individual students, along with other elements of the TVAAS approach, effectively purges the results of any bias that would otherwise arise as a result of the variation in student backgrounds (Lockwood and McCaffrey, 2007, p. 244.) At the same time, however, the mixed-methods approach cannot control for bias when the student population is stratified. A stratified student population is one “in which there are disjoint groups of students such that students within a group share teachers but students in different groups never share any teachers” (McCaffrey and Lockwood, 2007, p. 245).

Ballou, Sanders, and Wright (2004) reinforce these conclusions empirically in the context of the TVAAS model. To examine the effects of student level covariates, the authors add them to the TVAAS model in a two-stage approach. They begin with a first-stage equation in which student achievement gains are estimated as a function of student characteristics and standard teacher fixed effects (not the teacher fixed effects that emerge from the TVAAS model). The inclusion of the teacher fixed effects ensures that the estimated coefficients of the student characteristics are uncorrelated with any time invariant component of teacher quality. They then
use the estimated coefficients of the student characteristics to adjust the gain scores for each student and rerun the TVAAS model with the adjusted student gain scores. Consistent with the findings of Lockwood and McCaffrey (2007), the authors conclude that the use of the adjusted gain scores does not significantly change the estimates of teacher effects and hence that the unadjusted TVAAS model does an acceptable job of controlling for student-level covariates.

The results differ, however, when Ballou et al. (2004) make similar adjustments for contextual factors (such as the percent of students in a grade or school eligible for free and reduced price lunches). In that case, the TVAAS results change significantly, are implausibly large in some grades, and are sensitive to minor changes in model specification. Thus, consistent with the findings of Lockwood and McCaffrey, the stratification of students across schools renders the TVAAS model far less useful.

Although a discussion of the policy implications of these results is beyond the scope of this paper, it is worth highlighting that neither the value added approach nor the mixed methods approach to the estimation of teacher effects appears to generate sufficiently reliable and stable estimates of the causal effects of individual teachers for policymakers to use them for high stakes decisions about teachers. That need not rule out, however, the use of the results of value added modeling for lower stakes personnel decisions within a school. For ideas on the possibilities, see Rivkin (2007).

**Part 3**

**Are Teacher Credentials Predictive of Student Achievement?**

A common view, based on several early studies in the tradition of education production functions and well-publicized reviews of those studies by Eric Hanushek (1997), is that teacher credentials are not very predictive of student achievement, and hence are not useful as policy levers for improving schools. More recently, researchers have taken advantage of the richness of longitudinal administrative data to estimate the effects of credentials in a two-step procedure. In the first step, they estimate teacher effects using one of the approaches discussed in Part 2, and then in the second step, they explore the extent to which the variation in those effects can be explained by variation in teachers’ credentials. Because such researchers typically find little relation between teacher credentials and teacher effects, their findings reinforce the standard view that teacher credentials are not predictive of student achievement. The validity and usefulness of this two-stage approach, of course, depends in part on the validity of the estimated teacher effects.

Other researchers have taken advantage of the newly available rich administrative data to examine the predictive power of teacher credentials more directly. Their strategy is simply to replace the teacher variables in value added or gain models with a vector of teacher credentials. More recent work, including some by me and my Duke colleagues, has generated new and somewhat more positive results about the relationship between teacher credentials and student achievement. Because much of the recent literature has been reviewed elsewhere (see, for example,
Goldhaber, 2008), my discussion of the credentials literature is highly selective, is intended to be illustrative only, and draws heavily on my own research.

**ARE THE EFFECTS OF TEACHER CREDENTIALS BIG OR SMALL?**

Most value added or gain models that focus on teacher credentials are based on measures of student level test scores that have been normalized by year, grade, and subject. Thus the estimated coefficients that emerge from models explaining student test scores or gains in test scores are calibrated in terms of fractions of a standard deviation. For example, a standard finding in such models is that, relative to a teacher with no experience, the first year or two of experience is associated with an increase of student achievement of about 0.06 standard deviations, all other factors held constant. The apparently small size of this estimate, along with similar or smaller estimated coefficients for other teacher credentials, has led some people to argue that even if teacher credentials emerge as statistically significant determinants of student achievement, they may be inconsequential from a policy perspective. A coefficient of 0.06, for example, is tiny compared to black-white test score gaps that have historically ranged from 0.5 to 1.0 standard deviations and is well below the 0.20 effect size often deemed small or moderate in the education literature, and even further below the mean effect size of 0.33 standard deviations that emerged from a study of 61 random assignment studies of school interventions at the elementary level (Hill et al, 2007, cited by Boyd et al, in progress).

A new paper by Boyd, Grossman, Lankford, Loeb and Wyckoff (in press) argues that, if correctly interpreted, the effect sizes of teacher credentials are far larger than they first appear. The authors argue first that the coefficients of teacher credentials should be interpreted relative to the standard deviation of gain scores, not the standard deviation of test score levels. That argument follows directly from the cumulative nature of education. Given that estimates of teacher credentials in the context of a value-added or gains model are specifically designed to capture how the student’s fifth grade teacher, for example, affects the student’s gain in achievement during that year, it would not be appropriate to compare the estimated effect of the one-year teacher intervention to dispersion in the level of test scores, which reflect the cumulative effects of teacher and other variables over a longer period of time.

In addition, the authors argue that any interpretation of the estimated coefficients should account for the measurement error in the reported test score or scores. In particular, the coefficients should be compared to the dispersion in true achievement gains rather than in the measured achievement gains. This argument is consistent with the use of “reliable” variance discussed above in connection with research by Rowan et al. (2002). For any student, true achievement gains differ from actual achievement gains both because of the measurement error associated with the test itself and because the student may have been particularly alert or inert on the day(s) of the test. Failure to account for these measurement errors is particularly problematic when the focus is on achievement gains because gains are based on two test scores, both of which are measured with error. Importantly, the measurement error does not complicate the estimation of the effects of teacher credentials; it only affects the interpretation of the effect sizes. The elimination of measurement errors generates standard deviations in “true” achievement gains that are far smaller than those in measured achievement gains.
The authors develop their argument analytically and then illustrate its importance using a model designed to determine the relationship between the student achievement of fourth and fifth graders in New York City and the credentials of their teachers. The model is a gains model with student, grade and time fixed effects. Table 2 reports the results of their analysis for four of the teacher attributes included in the model: first year of experience (relative to no experience), not certified (relative to being certified), attended a competitive college (reference not clear), and a teacher math SAT score that is one standard deviation above the average (relative to a teacher with a mean SAT score). The estimated coefficients, three of which are statistically significant at the one percent level and one at the five percent level are shown in the first column. As is common in this literature these reported coefficients are expressed in terms of the observed variation in the test score level. In the second column these coefficients are measured relative to the standard deviation of the observed test score gain and in the third column relative to what the authors refer to as the gain in the universal (or “true”) test score.

### Table 2.
**Estimated Effect Sizes for Teacher Attributes Model for Math Grades 4 and 5 with Student Fixed Effects, NYC 2000-2005**

<table>
<thead>
<tr>
<th>Effect</th>
<th>S.D. of Observed Score</th>
<th>S.D. of Observed Gain Score</th>
<th>S.D. of Universal Score Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>First year of experience</td>
<td>0.065**</td>
<td>0.103</td>
<td>0.253</td>
</tr>
<tr>
<td>Not certified</td>
<td>-0.042**</td>
<td>-0.067</td>
<td>0.162</td>
</tr>
<tr>
<td>Attended competitive college</td>
<td>0.014*</td>
<td>0.022</td>
<td>0.054</td>
</tr>
<tr>
<td>One S.D. increase in math SAT score</td>
<td>0.041**</td>
<td>0.065</td>
<td>0.158</td>
</tr>
</tbody>
</table>

Source: Boyd, Grossman, Lankford, Loeb, and Wyckoff, in progress. Table from power point slides for presentation at AEFA meetings, Denver, April, 2008.

** 1% statistical significance  
* 5% statistical significance

The take-away point is that expressing the estimated coefficients relative to the standard deviation of the observed gains rather than to the standard deviation of the observed levels raises the effect sizes by about 50 percent and expressing them relative to the standard deviation of the “true” gain score raises them by about 400 percent. Thus, assuming this approach makes sense (which I believe it does), the estimated effect sizes are nontrivial. In particular, the effect of one year of teaching...
experience appears to raise achievement by an amount equal to 25 percent of a standard deviation of the gains in true achievement. Because this work on interpreting effect sizes is still in progress, it has not been fully vetted and some of the details of the estimates may change as the authors complete the research. I include it here because I believe that far more research would be useful along these lines to help both researchers and policymakers correctly interpret the magnitudes of the achievement effects that emerge from value added models of teacher credentials.

**STRATEGIES TO ESTIMATE THE ACHIEVEMENT EFFECTS OF TEACHER CREDENTIALS**

In a series of recent papers, my colleagues, Charles Clotfelter and Jacob Vigdor, and I have used a number of different strategies to estimate the effects of teacher credentials on student achievement. All the models are variations of the value added or gains models discussed in Section 2, and the research is based on rich administrative data from North Carolina accessed through the North Carolina Education Research Data Center. North Carolina is a particularly fitting state for this research because it has been administering end-of-grade tests for all students in Grades 3-8 and end-of-course tests in certain subjects for all students in high school since the early 1990s, and all the tests are closely linked to the statewide curriculum. In addition, since 1996/1997, teachers in schools that successfully raise student tests scores have been eligible for salary bonuses. As a result, teachers have an incentive to teach the material included on the state curriculum and students have an incentive to learn it.

Three of the papers (one of which is simply a longer version of a shorter published paper) focus on achievement at the elementary level and the fourth at the high school level. The main challenge we faced in this research was to devise credible ways of measuring the effects of teacher credentials given the nonrandom sorting of students and teachers among schools and across classrooms within schools.

**Clotfelter, Ladd and Vigdor (2006)**

The initial paper in this sequence was based on cross sectional data for one cohort of fifth grade students. We began by documenting the positive matching of students to teachers both across schools, and to a much lesser extent across classrooms within schools, where the term positive matching denotes that the more advantaged and higher performing students tend to have the teachers with the stronger credentials. Because the data were cross sectional, we were not able to use student fixed effects to address the expected upward bias in the estimates of teacher credentials.

Instead, we used three other strategies to minimize the bias. First, in addition to a standard set of student demographic variables, we added an extended set of student variables based on survey responses collected at the time the students were tested. These variables include information on time spent on homework, use of computers, and time spent watching television. Including these explanatory variables was helpful in that it reduced the magnitude of the error term in the student achievement equation, thereby reducing the room for reverse causation.

Second, we added school-level fixed effects, which was feasible because of our ability to match students to teachers at the classroom level. The inclusion of these
fixed effects meant that we were identifying the effects of teacher credentials based only on the variation across teachers within each school and thereby, we eliminated much of the statistical problem that emerges because of the sorting of teachers across schools.

Third, we addressed any remaining problems associated with the nonrandom assignment of students and teachers to classrooms within schools by restricting the analysis to the subsample of schools that, based on a number of observable characteristics, appeared to be assigning fifth grade students in a balanced way across classrooms, and hence to teachers within the school.

Finally, we tested the validity of our strategies by observing how the coefficients of the teacher credentials differed in specifications with and without the prior year achievement term. Our logic was that the closer our statistical solutions approached the gold standard of a random experiment, the smaller should be the role of the prior achievement term. In the context of a true random experiment, controls for baseline characteristics of students are not needed to generate unbiased estimates of treatment effects.

Clotfelter, Ladd, and Vigdor (2007a, b)
The first of these two papers is a shorter version of the second paper and focuses mainly on the results. The second paper provides the full models.

### Table 3.
**Model Descriptions Used by Clotfelter, Ladd, and Vigdor**

<table>
<thead>
<tr>
<th>Model</th>
<th>Dependent Variable (Achievement)</th>
<th>Lagged Achievement Included?</th>
<th>Type of Fixed Effect</th>
<th>Likely Direction of Bias of Effects of Teacher Credentials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>levels</td>
<td>yes</td>
<td>None</td>
<td>upward</td>
</tr>
<tr>
<td>Model 2</td>
<td>levels</td>
<td>yes</td>
<td>school</td>
<td>unclear, but small</td>
</tr>
<tr>
<td>Model 3</td>
<td>gains</td>
<td>no</td>
<td>school</td>
<td>downward</td>
</tr>
<tr>
<td>Model 4</td>
<td>levels</td>
<td>no</td>
<td>student</td>
<td>Downward</td>
</tr>
<tr>
<td>Model 5</td>
<td>gains</td>
<td>no</td>
<td>student</td>
<td>upward</td>
</tr>
</tbody>
</table>

A logical next step was to use longitudinal data for multiple cohorts of students to explore the same set of issues. The advantage of the longitudinal data set was that we were able to include student fixed effects in our models. In fact, though, we estimated five different models for third through fifth graders to explore the effects of different specifications, including three that did not include student fixed effects. We did so because we concluded that none of the models was capable of generating perfectly clean estimates of the effects of teacher credentials. Because testing does not start in North Carolina until grade three and it was not possible to identify a student’s specific teacher in math or reading after Grade 5, we restricted the analysis to student test scores in Grades 3-5. In models with prior year achievement, the models
refer to test score levels or gains for grades four and five. The short panels for each of our cohorts ruled out any instrumental variable strategy to counter the bias that arises from having the lagged achievement variable as an explanatory variable.

For each model, we predicted the direction of the bias based on conceptual and empirical considerations. Model 2 is quite similar to the model in our 2006 paper using cross-sectional data, but it is estimated with multiple cohorts of students. As elaborated in a conceptual note by Rivkin (2006) and shown in the bottom two rows of the Table 3 table, of particular interest is that the inclusion of student fixed effects is predicted to generate downward biased estimates of teacher credentials when the dependent variable is specified in levels and upward biased estimates when it is specified as an achievement gain. For reasons we spell out in the two papers, we prefer models 4 and 5, but with the recognition that model 4 provides a lower bound estimate of the effects of teacher credentials and model 5 an upper bound estimate.

Clotfelter, Ladd, and Vigdor (2007c)

Most of the research on teacher credentials, including the three papers just discussed, focuses on teachers at the elementary level. In this final paper, we shift the focus to the ninth and tenth grades and the relationship between student achievement on five courses (English 1, algebra I, geometry, biology, and economics, law and politics). This high school analysis is feasible in the North Carolina context because of the existence of statewide end-of-course tests in each of these subjects.

This paper makes a methodological contribution by its use of student fixed effects in the context of a model estimated across subjects rather than the more typical approach, over time. The inclusion of the student fixed effects means, as would be the case in longitudinal studies, that the effects of teacher credentials are estimated within students. In this case, they are based only on the variation in teacher credentials across the subjects for each student. This approach goes a long way in addressing the selection problem, provided students are assigned to classrooms on the basis of their overall, or average, ability or motivation, rather than on their likely success in a specific subject. Although we provide evidence in the paper in support of this key assumption, we cannot completely rule out the possibility that subject-specific unmeasurable abilities of students may be correlated with the teachers to whom they are assigned. This concern is analogous to the concern that arises in the context of longitudinal models about the role of time-varying student characteristics in the assignment of students to classrooms.

Effects of Credentials: Differences between Elementary and High School Teachers

Tables 4-9 summarize the results for various teacher credentials for teachers in elementary schools and high school that emerge from the Clotfelter, Ladd, and Vigdor research. In all cases, the entries emerge from models that include all the other teacher credentials, student fixed effects and a variety of other covariates. The tables are designed to highlight both the similarities and differences in the estimates. The elementary school results come from model 4 described above, which is based on longitudinal data with student fixed effects. The reported results should be interpreted as lower bound estimates.
Consider first the results for teacher experience in Table 4. Consistent with other studies, two-thirds of the impact of teacher experience (and about 5/6 at the high school level) is associated with the first few years of experience. In addition, as is quite common in past research, the estimates are larger for math than for reading at the elementary level. The subsequent tables report elementary school results only for math; in all cases the unreported comparable coefficients for reading are somewhat smaller. Further, the similarity in the results for teacher experience for the first two years for elementary and high school teachers in North Carolina and also with the 0.06 estimate for New York City reported in Table 2 is striking given the different methods and data used.

Table 4: Achievement Effects of Teacher Credentials: Teacher Experience

<table>
<thead>
<tr>
<th>Years of Exp. (base = 0)</th>
<th>Elementary</th>
<th>High School</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Math</td>
<td>Reading</td>
</tr>
<tr>
<td>1-2</td>
<td>0.057**</td>
<td>0.032**</td>
</tr>
<tr>
<td>3-5</td>
<td>0.072**</td>
<td>0.046**</td>
</tr>
<tr>
<td>6-12</td>
<td>0.079**</td>
<td>0.053**</td>
</tr>
<tr>
<td>13-20</td>
<td>0.082**</td>
<td>0.062**</td>
</tr>
<tr>
<td>21-27</td>
<td>0.092**</td>
<td>0.067**</td>
</tr>
<tr>
<td>&gt;27</td>
<td>0.084**</td>
<td>0.062**</td>
</tr>
</tbody>
</table>

Source: Elementary results are from Clotfelter, Ladd and Vigdor 2007b, Tables 2 and 3. High school results are from Clotfelter, Ladd and Vigdor, 2007c (revised 2009).

** denotes statistical significance at the 0.01 level; * at the 0.05 level.

Subsequent tables for other credentials also show remarkable similarities between the elementary and high school results. At both levels, nonregular licensure, including lateral entry licenses, are negatively associated with student achievement relative to regular licensure, with the effects somewhat larger at the high school level (Table 5). Also at both levels, it appears that teachers who subsequently are National Board Certified, are more effective than other teachers, a finding that provides support for the view that the Board Certification process identifies the more effective teachers. But, in contrast to the elementary level, at the high school level it appears that Certification itself may be associated with higher student achievement, leading to what some have called a positive human capital effect of the process (Table 6). With respect to master’s degrees, it appears that elementary school teachers who invest in a master’s degree part way into their teaching career are somewhat less effective than other teachers. At the high school level, in contrast, such master’s degrees are predictive of small positive effects on student achievement (Table 7). We find this difference in results quite plausible given that master’s degrees for high school teachers are likely to be more closely related to the...
subject taught. In neither case, however, do we separate the causal effect of getting a master’s degree from the selection effect of the decision to get one.

### Table 5:
**Achievement Effects of Teacher Credentials:**
**Licensure Status (comparisons are to regular licensure)**

<table>
<thead>
<tr>
<th></th>
<th>Elementary Math (lower bound)</th>
<th>High School</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lateral entry</td>
<td>-0.033*</td>
<td>-0.057**</td>
</tr>
<tr>
<td>Other</td>
<td>-0.033**</td>
<td>-0.074**</td>
</tr>
</tbody>
</table>

Source: Elementary results are from Clotfelter, Ladd and Vigdor 2007b, Table 2. High school results are from Clotfelter, Ladd and Vigdor, 2007c (revised 2009).

* statistical significance at the 0.05 level.

** statistical significance at the 0.01 level.

### Table 6:
**Achievement Effects of Teacher Credentials:**
**National Board Certification**

<table>
<thead>
<tr>
<th></th>
<th>Elementary Math (lower bound)</th>
<th>High School</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBCT -2</td>
<td>0.024**</td>
<td>N.A.</td>
</tr>
<tr>
<td>NBCT -1</td>
<td>0.018**</td>
<td>N.A.</td>
</tr>
<tr>
<td>Pre NBCT</td>
<td>N.A.</td>
<td>0.022**</td>
</tr>
<tr>
<td>NBCT current</td>
<td>0.018**</td>
<td>0.049**</td>
</tr>
<tr>
<td>NBCT has</td>
<td>0.022**</td>
<td>0.049**</td>
</tr>
</tbody>
</table>

Source: Elementary results are from Clotfelter, Ladd and Vigdor 2007b, Table 6. High school results are from Clotfelter, Ladd and Vigdor, 2007c (revised 2009). NBCT-2 denotes two years prior to the certification year, NBCT-1 denotes one year prior to the Certification year. Pre NBCT notes any year prior to certification. N.A. indicates not applicable because the variable was not included.

* statistical significance at the 0.05 level.

** statistical significance at the 0.01 level.

As shown in Table 8, the average predictive effects of teacher licensure tests are somewhat smaller at the high school level than at the elementary level. At the high school level, we are also able to look at the relationship between subject-specific teacher test scores and student performance by subject. Emerging from that analysis is the finding that higher teacher licensure test scores in math are clearly predictive of higher student achievement in algebra and geometry. Similarly, higher teacher test scores in biology predict higher achievement in biology, but the
Coefficient is smaller than for math. Contrary to our expectations, students in English do less well, all else held constant, when they have teachers with higher licensure test scores in English. Finally, in table 9, we report results by certification status, but only at the high school level. The results indicate that certification in math and English are highly predictive of student achievement in those subjects and the effects are large. No teacher certification effects emerge for biology and ELP.

Table 7:
Achievement Effects of Teacher Credentials:
Master’s Degrees

<table>
<thead>
<tr>
<th></th>
<th>Elementary Math (lower bound)</th>
<th>High School</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA before teaching</td>
<td>-0.001</td>
<td>-0.006</td>
</tr>
<tr>
<td>MA 1-5 years into teaching</td>
<td>0.004</td>
<td>0.009**</td>
</tr>
<tr>
<td>MA degree 5+ years into teaching</td>
<td>-0.010 **</td>
<td>0.009 **</td>
</tr>
</tbody>
</table>

Source: Elementary results are from Clotfelter, Ladd and Vigdor 2007b, Table 5. High school results are from Clotfelter, Ladd and Vigdor, 2007c (revised 2009).

** statistical significance at the 0.01 level
* statistical significance at the 0.05 level

Table 8:
Achievement Effects of Teacher Credentials:
Teacher Test Scores

<table>
<thead>
<tr>
<th></th>
<th>Elementary Math (lower bound)</th>
<th>High School</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher test score</td>
<td>0.011**</td>
<td>0.007**</td>
</tr>
<tr>
<td>Math score for alg. and geo.</td>
<td>0.047**</td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td>0.016*</td>
<td></td>
</tr>
<tr>
<td>ELP</td>
<td>0.001**</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>-0.024 ** unexp</td>
<td></td>
</tr>
</tbody>
</table>

Source: Elementary results are from Clotfelter, Ladd and Vigdor 2007b, Table 3. High school results are from Clotfelter, Ladd and Vigdor, 2007c (revised 2009), Table 4.

Unexp. Denotes that the sign of the coefficient was unexpected.

** statistical significance at the 0.01 level
* statistical significance at the 0.05 level
Table 9: 
Achievement Effects of Teacher Credentials: 
Teacher Certification, High School Only

<table>
<thead>
<tr>
<th>Subject</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alg. and geo.</td>
<td>Teacher certified in math 0.103**</td>
</tr>
<tr>
<td>Biology</td>
<td>Certified in biology -0.016</td>
</tr>
<tr>
<td>ELP</td>
<td>Certified in ELP 0.004</td>
</tr>
<tr>
<td>English</td>
<td>Certified in English 0.0113**</td>
</tr>
</tbody>
</table>

Source: Clotfelter, Ladd and Vigdor, 2007c (revised 2009), Table 4.
** statistical significance at the 0.01 level

Magnitudes of the Effects

To the extent that Boyd et al. (in progress) are correct that the estimated effect sizes should be adjusted upward to take account of measurement error, these North Carolina estimates, like those for New York City, understate the relevant effect sizes. We have also used other methods for evaluating the effect sizes. In our work on elementary school teachers, we compare the predicted effects of teachers with “strong” bundles of credentials with those of teachers with “weak” bundles and conclude that the effects of teacher credentials are sufficiently large to offset the estimated effects of low parental education on student achievement for math but not fully for reading. In addition, we conclude that teacher credentials are far more predictive of student achievement than class size reductions of moderate size.

At the high school level, we conclude that a teacher at the 90th percentile of the predicted distribution of the achievement based on teacher credentials would increase student achievement by about 0.18 standard deviations relative to a teacher at the 10th percentile. Thus, we conclude that teacher credentials are important predictors of student achievement. At the same time, however, we are careful to note that teachers with similar bundles of credentials exhibit substantial variation in their effectiveness.
REFERENCES


ENDNOTES

1 This lower bound estimate exceeds the comparable estimate of 0.11 standard deviations reported in Rivkin, Hanushek, and Kain (2005), but the results are not directly comparable. Although this other study also refers to within-school differences, it focuses on grade level differences from one year to the next. Another important difference is that this other study is based on raw gains, not the standardized gains used for the results in Table 1. The standard deviation of the raw gains is about two-thirds of the standard deviation of the standardized gains. Putting the two estimates on the same scale would increase the 0.11 estimate to 0.15 standard deviations. (Hanushek, Kain, O’Brien, and Rivkin, 2005, p. 14).

2 This discussion is based primarily on the discussion of McCaffrey, Lockwood, Koretz, and Hamilton (2003), pp. 24-30.

3 Often included as time varying student variable are indicators for whether a student has changed schools, either independent of other students or as part of a move with others from one level of schooling to another.

4 As discussed in Lockwood, McCaffrey and Sass, 2008, it is crucial who the reference, or left-out, teacher is in the regression model. If the reference teacher is far from the mean in terms effectiveness the estimates of the standard errors associated with the individual teacher effects are not plausible. Instead, it turns out to be best to specify an average teacher as the reference teacher.

5 At the elementary level, the nonrandom matching of students to teachers appears to be a far larger problem than the nonrandom matching of students to teachers across classrooms within schools in North Carolina (Clotfelter, Ladd and Vigdor, 2006).

6 Another potential explanation for the findings relates to the linearity of the model. In current work, Austin Nichols at the Urban Institute is currently testing that possibility.

7 See McCaffrey et al, 56-62.

8 Also included in the model are whether the student changed schools between years, classroom variables that include the proportion of students who are black or Latino, the proportion who receive free-or reduced price school lunch, class size, the average number of student absences in the prior year, the average achievement scores of students in the prior year, and the standard deviation of student test scores in the prior year. Teaching experience is measured by separate indicator variables for each year of teaching experience up to a category of 21 or more years. Other teacher qualifications include whether the teacher passed the general knowledge portion of the certification exam on the first attempt, certification test scores, whether and in what area the teacher was certified, the Barron’s ranking of the teacher’s undergraduate college, math and verbal SAT scores, the initial path through which the teacher entered teaching, e.g., a traditional college recommended program or the New York City Teaching Fellows program, and an interaction term of the teacher’s certification exam score and the portion of the class eligible for free lunch. The standard errors are clustered at the teacher level to account for multiple student observations per teacher.

9 None of the models from which these results emerge include teacher fixed effects. Hence it is not possible to determine from this table alone whether the patterns of the estimates over time reflect learning on the job or the patterns of teacher attritions, either out of the elementary schools or out of the core ninth and tenth grade courses at the high school level. This issue is discussed in more detail and with additional estimates in both the cited papers.