## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Teacher Effects: What Do We Know?</td>
<td>3</td>
</tr>
<tr>
<td><em>Helen F. Ladd</em></td>
<td></td>
</tr>
<tr>
<td>Teacher Labor Markets and Teacher Labor Market Research</td>
<td>27</td>
</tr>
<tr>
<td><em>Susanna Loeb and Tara Béteille</em></td>
<td></td>
</tr>
<tr>
<td>Using Standardized Observation to Conceptualize, Measure, and Improve Classroom Processes</td>
<td>59</td>
</tr>
<tr>
<td><em>Robert C. Pianta and Bridget K. Hamre</em></td>
<td></td>
</tr>
<tr>
<td>Contributors</td>
<td>79</td>
</tr>
</tbody>
</table>
Teacher quality has always been a salient issue for policymakers, practitioners, and researchers. The advent of No Child Left Behind and other accountability policies has accelerated interest in teacher quality issues by pushing them front and center in state and local school district policy conversations. In an attempt to deepen and broaden the debate on teacher quality, we commissioned three papers by leading education policy experts and invited the authors to present and discuss the papers at conference held on May 1, 2008. To hasten their dissemination, we have opted to publish the papers online.

Do Teachers Matter?

In the first paper, Helen Ladd examines what we know and what we need to learn about teacher effects. Her first question is, do teachers matter? As might be expected, her literature review shows that researchers have a range of answers to this question, from “somewhat” to “quite a bit.” A paper by Nye et al. (2004) illustrates the size of these effects by taking advantage of data on teachers who were randomly assigned to their students. Their calculations suggest that the impact of having a teacher at the 25th as opposed to the 75th percentile of the effectiveness distribution produces one-half a standard deviation improvement in math and about one-third a standard deviation advantage in reading. These are large differences; effects this size imply that having an effective as opposed to ineffective teacher for several years would eliminate much of the racial and ethnic differences in the achievement gap.

Ladd’s second question flows from the first: given that teachers matter, how can we identify effective teachers? One obvious candidate is a teacher’s credentials: do teachers with more experience, more teaching degrees, or who graduated from better colleges teach better? Answers here are even more controversial, but emerging work suggests larger achievement impacts from teacher experience and credentials than had been previously thought.
Teacher Quality:  
Broadening and Deepening the Debate

What Happens Inside the Classroom?

At best, the literature reviewed by Ladd treats teacher quality as a black box, revealed in part by a teacher’s ability to boost student test scores or correlated with his or her personal characteristics. But not addressed are how, exactly, the test score gains come about, or why experience or credentials matter. In the second paper, Robert Pianta and Bridget Hamre address teacher quality issues with a systematic examination of what actually goes on in effective and ineffective classrooms. Are students engaged in high-quality, interactive instructional activities with a responsive teacher? Or are most spending their time filling out worksheets and watching their teacher struggle to keep ill-behaved students under control? These key dimensions can be measured systematically, Pianta and Hamre argue, and envision a time when such measurement can be used to monitor and improve the quality of what actually goes on in classrooms.

Who Becomes a Teacher?

In the third paper, Susanna Loeb and Tara Béteille step back from individual teachers and their classrooms to take a more macro view of teacher labor markets and their implications for student learning. On average, students have older, more experienced and better credentialed teachers than before, a beneficial development to the extent that these qualities help promote student learning. On the other hand, shattered glass ceilings have opened up many professional jobs for women, leaving fewer women from the top ranks of their high schools to enter the teaching profession. These averages conceal enormous variability in teacher quality between rich and poor school districts, with the differences perpetuated in part by the preferences of teachers to teach in school districts that are close to where they grew up. These and a host of other features of teacher labor markets help define the promise and challenges of efforts to increase the quality of the teacher work force.

* * *

These three chapters document how much we have learned in the last decade about the nature and impact of a high-quality teacher. At the same time, pressing policy concerns push us toward ever more concrete policy recommendations for ensuring that all children enjoy the many benefits from high-quality teachers.

The papers and conference were sponsored by Northwestern University’s Multidisciplinary Program in the Education Sciences. Deborah Ball, Henry Braun, Tom Cook, Larry Hedges, and Derek Neal served as discussants for the papers. At the conference, Tom Kane presented his previously published proposal for improving teacher quality by basing teacher tenure decisions, in part, on past test-score gains. Interested readers can access this paper here: http://www.brookings.edu/views/papers/200604hamilton_1.pdf. We are grateful to the Institute of Education Sciences for its generous funding of the program. We are also grateful to Hannah Wallerstein, Cindy Sigal, Michelle Reininger, Larry Hedges, Lindsay Chase-Lansdale and David Uttal for their many contributions to the conference design and logistics.
Teacher Effects: What Do We Know?

Helen F. Ladd

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.
**Classroom and Teacher Differences in Student Achievement Gains**

<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:

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

\[ \text{[equation 1]} \]

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
- \( \varepsilon_{it} \) 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 β be the effect of T and α the rate at which knowledge persists, we can rewrite equation 1 as:

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

and, after rearranging terms, as:

\[ A_{it} = \beta T_{it} + \alpha (\beta T_{it-1} + \alpha \beta T_{it-2} + \alpha^2 \beta T_{it-3} + \ldots ) + \epsilon_{it} \]  
[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:

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

where the error term, \( u_{it} \), is equal to \( \epsilon_{it} - \alpha \epsilon_{it-1} \).

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:

\[ A_{it} - A_{it-1} = \beta T_{it} + \epsilon_{it} \]  
[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.

\[
A_{it} = \alpha A_{i,t-1} + \beta T_{it} + \gamma X_{it} + \delta C_{it} + \theta_i + \eta_{it} \quad \text{[equation 6]}
\]

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, $\beta_t$ the true value and $\varepsilon$ 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:

$$\beta_{t,*} = \beta_t + \varepsilon$$

[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:

$$\lambda \beta_{t,*} + (1-\lambda) \text{mean } B_{t,*}$$

[equation 8]

where:

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

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 Sizes: Estimated effects relative to</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>
<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.\textsuperscript{9} 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 0.01 level
* statistical significance at the 0.05 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 0.01 level
* statistical significance at the 0.05 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.: Teacher certified in math</td>
<td>0.103**</td>
</tr>
<tr>
<td>Biology: Certified in biology</td>
<td>-0.016</td>
</tr>
<tr>
<td>ELP: Certified in ELP</td>
<td>0.004</td>
</tr>
<tr>
<td>English: Certified in English</td>
<td>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


**Teacher Quality:**
*Broadening and Deepening the Debate*


**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.
The research on teacher labor markets is quite large and expanding; yet, as in most areas of education research, our knowledge is full of holes and only gets us a little ways towards identifying productive policy directions. As such, there is plenty of room for new research—describing labor market dynamics, developing and substantiating theories about the mechanisms driving the trends and relationships observed, developing instruments for measurement, and evaluating programs. This paper begins by providing an overview of what we know about teacher labor markets in the United States. It is not an exhaustive review of the literature, but rather a summary of the state of our knowledge with illustrative cites to findings in the literature. The final section identifies possible directions for new research. In doing so, it distinguishes supply-side and demand-side factors during four stages of a teacher’s career—pre-teaching, early-teaching, middle-teaching, and late-teaching.

**Describing the Teacher Workforce**

**THE TEACHING FORCE**

The number of public elementary and secondary school teachers in the United States has grown steadily over the last 50 years. In 1955, there were 1.14 million public elementary and secondary school teachers. The 2003-2004 Schools and Staffing Survey (SASS) puts this number at 3.25 million for the country’s 15,500 school districts, serving approximately 47.3 million students (NCES, 2006). The increase in teaching staff has been driven primarily by rising student enrolment and falling student-teacher ratios. Student enrolment increased in the 1950s and 1960s owing to the post-war baby boom. It declined by approximately 5 million between 1970 and 1990, but has been increasing since.2 Student-teacher ratios decreased from 26.9 in 1955 to 14.5 in 2003-2004 (NCES, 2006). Part of this decline

---


comes from class size reduction policies targeted at all students, but also from laws mandating the provision of education to all handicapped children (Hanushek and Rivkin, 2002). The special education sector has become more staff-intensive; from 194,802 special education teachers in 1978, the number of teachers in this sector rose to 307,575 in 1990 and 412,750 in 2003-2004 (Hanushek and Rivkin, 2002).

AGE AND EXPERIENCE

The average age of teachers has increased over the last 30 years. The median age of teachers was 41 in 1961, falling to 33 in 1976, but increasing thereafter. The average age of public school teachers was 42.5 years in 2003-2004 (NCES, 2006). At least two forces have been driving the increase in the average age of teachers. First, teachers hired to educate the children of the baby boom era have aged and are now reaching retirement. These teachers are likely to retire over the next 10-15 years, creating a demand for new teachers. Second, those entering teaching today are older than in the past. For example, more than 80 percent of new teachers in New York were under age 25 in 1970. By the mid-1980s, this had decreased to roughly 40 percent. It has continued to decline slowly ever since (Loeb and Reininger, 2004).

In keeping with the changing age distribution of teachers, experience levels have changed over time. Whereas the 1987-1988 SASS found that only 9.9 percent of all public school teachers had taught for three or fewer years, the 2003-2004 SASS found that 17.8 percent of all full-time public school teachers had teaching experience of three or fewer years. Roughly the same percentage of teachers had been teaching for 20 or more years in 2003-2004 as in 1987-1988. In terms of numbers, however, there were over 160,000 more teachers in 2003-2004 with teaching experience of 20 or more years than there were in 1987-1988 (NCES, 2006).

Multiple studies have estimated the effects of teaching experience on students’ learning, though few have looked at the effects of teachers’ age. Using data on New York City schools for grades 4-8, Boyd et al. (2006a) found that, on average, first and second year teachers did not add as much to student learning in English language arts (ELA) or math as more experienced teachers did. Gains accrued thereafter, but stopped being substantial after the fifth year (Boyd et al., 2006b). Using a 10-year panel from North Carolina, and focusing on students from grades 3, 4 and 5, Clotfelter et al. (2006) found that the more experienced a teacher was, the more student test scores in reading and math increased over the course of a year. Compared with a teacher with no experience, the benefits of experience rose continuously with experience peaking at 21-27 years of experience. They also found that more than one-half of the gain occurred during the first few years of teaching.

This is consistent with other studies in Texas (Hanushek, Kain and Rivkin, 2003) and New Jersey (Rockoff, 2004). The evidence from New Jersey, however, suggests

---

3 The Federal Law enacted through the Education for All Handicapped Children Act in 1975 makes the provision of educational services to all mentally and physically handicapped children compulsory.

4 http://nces.ed.gov/surveys/sass/das.asp
that the effect of experience may vary by subject matter. Using a panel from New Jersey, Rockoff (2004) found that the impact of teacher experience on student vocabulary achievement increased until the sixth year of teaching after which it flattened. The impact on reading comprehension, however, increased monotonically past the tenth year. The effect of teacher experience on math computation skills, on the other hand, increased till year three after which it began to decrease. All of these studies suggest that experience matters for student learning, but that on average, the gains to experience are greatest in the initial years of teaching.

The better performance of more experienced teachers could reflect either improvement with experience or the differential attrition of ineffective teachers. If those who are less effective, on average, are also the ones to leave initially, then what looks like gains to experience might simply be gains to more effective teachers regardless of experience. In other words, we might see more experienced teachers, on average, registering higher student test scores, even if they as individuals did not get any better with experience. Using a Florida Panel for grades 3-10, Harris and Sass (2006) find that while experience generated positive effects for student learning in both math and reading, those effects became very small when teacher fixed effects were included. This suggests it may be differential attrition, not improvement in teaching skills, which drive the better performance of more experienced teachers. Clotfelter et al. (2006), however, find little support for the differential attrition hypothesis from their analysis of North Carolina schools. They argue that positive returns to experience in their models come primarily from experience and not from a sample biased by the attrition of ineffective teachers. Thus, although it is clear that, on average, more experienced teachers are more effective than first-year teachers, the extent to which this is driven by learning or attrition is less clear. There are likely to be differences in learning opportunities available to teachers from place to place, which could influence the effect of teacher experience on student learning, as well.

GENDER

Approximately 75 percent of public school teachers are female, with 83.8 percent female in elementary schools and 57.3 percent in secondary schools (NCES, 2006). These proportions are similar across the urban and rural spectrum, although schools in the South and Midwest employ relatively more women than other schools (Bacolod, 2005). The proportion of female teachers has not changed dramatically over the last 50 years. Two things, however, have changed. First, the number of women completing college has risen dramatically. As a result, the test score of the average college graduate is now lower relative to the full distribution of high school students in a given cohort, than was previously the case, when a smaller proportion of high school students went onto college. Second, as Corcoran, Schwab, and Evans (2004) note, the labor market for women has changed considerably since the mid-1960s, with traditionally male professions such as law and medicine becoming increasingly open to women. Using data from five longitudinal surveys of high-school graduates spanning the classes of 1957 to 1992,
they found that while the math and verbal test scores of the average new teacher had fallen only slightly, the likelihood that a female from the top of her class would enter teaching had fallen dramatically. Bacolod (2005) reaches similar conclusions. Using indices of teacher quality such as test scores and selectivity of undergraduate institution, she establishes an empirical link between an increase in professional opportunities for women and a decline in the quality of teachers as measured by these indices.6,7

The research literature assessing the effect of teachers’ gender on student outcomes is relatively small. Ehrenberg, Goldhaber, and Brewer (1995) did not find a systematic relationship between teacher gender and student outcomes; while Nixon and Robinson (1999) found no relationship between gender and outcomes for boys but found that girls attending high schools with a higher proportion of female teachers had higher educational attainment. In a recent study of the relationship between teacher gender and student outcomes, Dee (2007) finds that same gender matches between teachers and students improves student learning. In particular, boys appear to learn less with female teachers.

TEACHER AND STUDENT RACE

The racial and ethnic makeup of teachers does not reflect that of their students in most school districts. The share of nonwhite students is much larger than the share of nonwhite teachers. For instance, the proportion of African-American and Hispanic students (16.8 and 17.7 percent, respectively) is nearly three times the percentage of African-American and Hispanic teachers (7.9 and 6.2 percent).

The underrepresentation of racial and ethnic minority teachers stems largely from their under-representation in the college-educated population. Among college graduates in 1976-1977, for example, 90 percent were white, 7 percent were African-American, and 2 percent were Hispanic. By 1999-2000, the gap had decreased slightly to 78 percent, 9 percent, and 6 percent respectively. Nevertheless, non-Hispanic white teachers were considerably overrepresented in the group of college graduates.

How important is having a teacher of the same race for student achievement? A recent study using data from the Tennessee STAR experiment in which students and teachers were randomly assigned to each other found that an additional year with an own-race teacher increased student performance by two to four percentile points (Dee, 2004). As Dee notes, a comparison with other estimated effects suggests these gains are considerable. Specifically, they are comparable to those associated with a small-class assignment. The results are in tune with those from

---

6 Bacolod points out that the results from her analysis on selectivity of undergraduate institution need to be interpreted in the light of two facts: (1) during the period under study, many universities curtailed undergraduate education programs, and (2) there is no empirically verified one-to-one link between majoring in education and becoming a teacher. See Bacolod (2005) for more.

7 While changes in female labor markets appear to be the major source of the decline in highly qualified women entering teaching, Bacolod points out the potential for additional explanations. For instance, women’s admission to professional programs, their increased access to credit markets for loans to pursue skill acquisition and even access to the pill, as well as unionization in teaching and deunionization in non-teaching, and the general rise in skill returns might also explain the above pattern. See Bacolod (2005) for more.
a large school district in Texas, where black students’ scores improved by 0.1 standard deviations when they had a black teacher compared with when they had a white teacher (Hanushek, Kain, O’Brien, and Rivkin, 2006). These results should be interpreted with caution to the extent that teacher quality varies systematically with school-level student racial composition, making it difficult to separate teacher quality from teacher race. If, for instance, the best white teachers self-select themselves into more affluent schools, leaving the least competent white teachers in schools with a high share of low-income, low-achievement black students, then such studies might end up comparing the “average” black teacher with a set of “below-average” white teachers, leading one to overstate the benefit of having a same-race teacher (Jacob, 2007).

EDUCATIONAL ATTAINMENT

Almost all public school teachers have bachelor’s degrees and nearly 41 percent have master’s degrees as their highest degree earned (NCES, 2006). In 1961, 15 percent of teachers did not have a bachelor’s degree, but by the early 1980s, nearly all teachers had completed an undergraduate degree. As an example, in 2003-2004, only 1.1 percent of all public school teachers did not have a bachelor’s degree (NCES, 2006). The percentage of teachers with master’s degrees as their highest degree has risen considerably, from approximately 23 percent in 1961 to 41 percent in 2003-04 (NCES, 2006). Degree attainment varies by the grade the teacher teaches, with high school teachers more likely to hold a master’s degree than middle school teachers, who in turn are more likely to hold a master’s degree than primary school teachers. However, there is little difference across community types (rural, suburban, and urban) in the percentage of teachers with masters’ degrees.

The increase in master’s degree attainment is, at least in part, related to changes in state requirements and the additional pay linked to educational attainment in district or state salary schedules. The incentives that encourage teachers to get a master’s, unfortunately, are not likely to have benefited students. Master’s degrees have not been found to predict higher student achievement, except for content specific masters’ degrees in high school mathematics. For example, using North Carolina data, Clotfelter et al. (2006) found no impact of master’s degrees on student achievement in elementary school; in some cases, the impact was negative, though they find more positive effects in high schools (Clotfelter et al., 2006). A study using Florida panel data also found that advanced degrees were not effective, on average, in increasing teacher productivity. There is some evidence that students of teachers with subject-specific master’s degrees learned more over the course of a year, but, as noted, this was only the case for high school mathematics, and has yet to be confirmed using current empirical techniques (Harris and Sass, 2006). Even here, it is unclear whether it is the master’s degree per se or greater interest in math (which presumably led them to the master’s degree) that leads to better student performance. If it is the latter, then these teachers might have helped improve student performance even without the master’s.
SUBJECT-MATTER KNOWLEDGE

Basic reasoning would lead one to expect teacher effectiveness to be linked to adequate subject-matter knowledge. There are several ways of measuring a teacher’s subject-matter knowledge—for instance, scores in field-specific examinations such as the Praxis series, teaching certificates, or undergraduate or graduate course taking. Although none of these captures subject-matter knowledge completely, they nevertheless give us some sense on average of a teacher’s content knowledge.

Most teachers have a graduate or undergraduate major or minor in their primary teaching field, and this has been increasing over the years (Ingersoll, 2003). As of 1997-1998, 86 percent English teachers in grades 7 through 12, 89 percent of social science teachers, 82 percent of math teachers, and 88 percent of science teachers reported having an undergraduate or graduate major or minor in their main teaching assignment field. The types of majors teachers have vary substantially by school level. High school teachers are far more likely to have degrees in traditional academic fields such as math or history (66 percent) than are middle school teachers (44 percent) or elementary school teachers (22 percent) (Loeb and Reininger, 2004).

Many teachers, however, also teach classes outside their primary teaching assignment—and they are much less likely to hold a major or minor in these areas. In 1999-2000, Ingersoll (2003) found that 38 percent of all grade 7-12 teachers who taught one or more math classes did not have either a major or a minor in math, math education, or related disciplines like engineering, statistics or physics. One-third of all 7-12th grade teachers who taught one or more English classes had neither a major nor minor in English or related subjects such as literature, communications, speech, journalism, English education, or reading education. In science and social studies, the numbers were slightly lower. Approximately 28 percent of all 7-12th teachers who taught one or more science classes lacked even a minor in one of the sciences or in science education. Finally, approximately 25 percent of those who taught one or more social studies classes lacked a minor in any of the social sciences, in public affairs, in social studies education, or in history (Ingersoll, 2003).

The No Child Left Behind (NCLB) Act of 2001 mandated that every student be taught by a “highly qualified” teacher by 2006. NCLB defines a highly qualified teacher as a fully state-certified teacher who holds a bachelor’s degree and demonstrates competency in the core academic subject or subjects he or she teaches. Under these standards, to be fully state-certified, a teacher must obtain a certificate appropriate to his or her level of experience and must not be in a position where certification or licensure requirements are waived on an emergency, temporary, or provisional basis. The law provides states considerable flexibility in determining the exact criteria for certification within the broad framework laid out. States, for instance, are allowed to determine their own requirements for indicators of subject-matter competence. Twenty-five states require high-school teachers to have a major in their primary subject area and to have passed a subject-matter exam. Six states require high school teachers to only have an undergraduate major

---

8 The law defines core academic subjects as the following: English, reading or language arts, mathematics, science, foreign language, civics and government, economics, arts, history, and geography (PL 107-110 §9101(11)). “Arts” as a subject was not further defined in the law. The Department’s guidance directed each State to determine its own definition (U.S. Department of Education, Dec. 2002, p. 14).
in the area, while eighteen other states require teachers to only pass a subject-
matter test in their primary teaching field (Boyd et al., 2006a). Note, however, that
there is considerable variation in the level of knowledge that constitutes a major or
that which is necessary for certification exams.

Although there is much rhetoric around the importance of subject-matter compe-
tence on teacher effectiveness, to date most research shows no strong relationship
between teachers’ subject matter knowledge and student test-score gains. For ex-
ample, in their study of New York City schools, Boyd et al. (2006b) find no relation-
ship between teacher’s undergraduate degree and student performance. Similarly,
using data from the San Diego Unified School District, Betts et al. (2003) find
no clear link between a student’s rate of learning at the elementary level and the
number of college courses completed by his or her teacher in a particular subject.
This should not be taken as evidence that content knowledge is not important, but
simply that it may not be the factor that most differentiates teachers’ effectiveness
in the classroom, especially in the early grades.

There is some recent evidence that suggests it is not content knowledge per se, but
pedagogical knowledge that is important for student learning. Ball, Rowan, and Hill
(2005) find that pedagogical knowledge for teaching math is significantly associat-
ed with student achievement for in grades 1 and 3, after controlling for key student
and teacher-level covariates. In high school, however, recent literature finds that
even the more general measures of teacher content knowledge might be associated
with learning. For instance, Clotfelter, Ladd, and Vigdor (2007) find some evidence
that teachers who obtained a masters’ degree while teaching add more value to
student learning in high school than do teachers without masters’ degrees.

**TEACHER ABILITY: TEST SCORES AND SELECTIVITY
OF UNDERGRADUATE INSTITUTIONS**

Although there is little evidence on the importance of content knowledge for
student learning, there is some evidence that teachers with greater general knowl-
edge and academic ability are more effective in the classroom. This relationship,
however, is not strong.

Teachers, on average, score below the typical college graduate on standardized
aptitude tests (Bacolod, 2005; Corcoran, Schwab, and Evans, 2004; Hanushek and
Pace, 1995). Focusing on the average alone, however, masks the fact that many
teachers score well on standardized aptitude tests. In a study of more than 300,000
prospective teachers who took a Praxis test between 1994 and 1997, Gitomer,
Latham, and Ziomek (1999) found that prospective teachers in academic subject
areas had SAT/ACT scores that were comparable, if not better, than the larger col-
lege graduate population. At the same time, those seeking licenses in non-academic
fields such as elementary education had much lower scores. The academic ability
of teachers has also changed over time. More than 20 percent of young female
teachers in the 1960s scored in the top 10 percent of their high-school graduating
cohort. By 2000, this number had dropped to 11 percent, although the changes in
other parts of the achievement distribution were not great (Corcoran et al., 2004).
For men in the top two decile groups, the drop in the probability of entering teach-
ing was comparatively lower. While 6.3 percent of men in deciles 9 and 10 of their
high school graduating cohort entered teaching in 1964, this figure had dropped to
3.8 percent in 2000. Bacolod’s findings complement this; she shows that among those with higher test scores, the predicted probability of entering alternative professions has increased dramatically.

Students of teachers with higher test scores tend to learn slightly more as measured by test score performance than other students. The relationship appears stronger in math than in reading. Using North Carolina data, Clotfelter et al. (2006) find that teachers who had scored two or more standard deviations above the average boosted students test scores by 0.068 standard deviations while those who scored two or more below the average reduced achievement gains by 0.062 standard deviations. They conclude that having a teacher at either extreme of the test score distribution has a far bigger effect on student math achievement than having an average teacher. Using data on New York City school teachers, Boyd et al. (2007b) find that teachers who passed the Liberal Arts and Sciences Test (LAST) state teacher-certification exam on their first attempt produced higher student math achievement than those who did not. They find no effects for student ELA performance. Importantly, however, they find that higher-scoring teachers on average have a greater effect on students with higher prior test scores. When teaching students with lower prior test scores, they tended to do no better, and in some cases worse, than lower scoring teachers.

VARIATION IN TEACHER CHARACTERISTICS ACROSS SCHOOLS

The description of the teacher workforce above masks the substantial variation in teacher characteristics across schools and school districts. Nationwide, schools with the highest minority enrolment, largest low-income enrolments and the most academically struggling students are also the ones most likely to have teachers with the weakest qualifications.

Certain features of the distribution of teachers stand out. First, there is greater variation in teacher credentials within individual cities than across cities. For instance, there are larger variations in teacher credentials, such as selectivity of undergraduate institution and average experience, among the schools in the Phoenix metropolitan area, than there are between the metropolitan areas of Phoenix and Detroit (Loeb and Page, 2001).

---

9 Corcoran, Schwab and Evans (2004) point out that these results should be interpreted with caution since the sample for men was much smaller than that for women. Also, male teachers are much more likely to be secondary school teachers.

10 See Bacolod (2005). As mentioned previously, Bacolod notes certain problems in using data on the selectivity of undergraduate institutions. First, during the period under study, many institutions curtailed undergraduate education programs. Second, there is no one to one link between education majors actually becoming teachers.

11 From the early 1960s through the mid 1990s, all elementary school teachers in North Carolina were required to take either the Elementary Education or Early Childhood Education test. The former included material on curriculum, instruction and assessment. Starting in the mid 1990s, teachers were required to take both that basic elementary test and one focusing on content. See Clotfelter et al (2006) for more.
This variation across schools within metropolitan areas is systematic. Schools with high minority enrolments also have higher proportions of teachers in their first three years of teaching, higher proportions of teachers with less than ten years experience, and the lowest proportion of teachers with more than twenty years experience. They also have the lowest share of teachers with certification in their primary or secondary teaching assignment. In the New York City school district, for example, there are large differences in teacher characteristics across racial and income groups (see Loeb and Reininger, 2004). As of 2000, 21 percent of nonwhite students had teachers who were not certified in any subject taught compared with only 15 percent of white students. Twenty-six percent of nonwhite students had teachers who failed the general knowledge certification exam compared with 16 percent of white students. Similarly, 22 percent of low-income students had teachers who were not certified in any subject they taught compared with 17 percent of higher income students. Thirty percent of low-income students had teachers who failed the certification exam, compared with 21 percent of higher income students.

There is also some variation across community type. Approximately 50 percent of all schoolteachers work in suburban settings, with the other half evenly distributed between rural and urban areas. Teachers in these settings are similar in terms of gender, experience and certification, yet fairly different when it comes to race, age and educational attainment (see Loeb and Reininger, 2004). Not surprisingly, nonwhite teachers more frequently teach in central cities than in urban fringe/large towns or rural/small towns (NCES, 2006). Fewer rural teachers hold master’s degrees compared with teachers in urban and suburban settings (NCES, 2006).

The choices individual teachers make with regard to job posting are influenced by multiple measured and unmeasured factors. The research literature has identified and assessed a number of these including wages and benefits, working conditions, entry requirements, and school location. These are believed to affect the supply decisions of teachers. They tell us whether college graduates will choose teaching as a profession, and if they make this choice, where they are likely to teach. The eventual outcome is, however, also influenced by factors originating from the school system, i.e. from those who demand teachers. Important among these factors are district hiring practices, contracts and bureaucratic features. We look at each in turn.

**WAGES**

Much research suggests that teachers are more likely to choose teaching when starting wages are high relative to wages in other occupations (see Bacolod, 2005; Corcoran et al., 2004; Hanushek and Rivkin, 2006). Drawing on multiple data sources, Bacolod (2005) finds that highly qualified teachers are particularly sensitive to changes in relative wages. The lower teachers are paid relative to professionals, the less likely high-quality educated women are to choose teaching (Corcoran et al., 2004). Approximately 16.5 percent of public school teachers who decided to move to another school between 2003-2004 and 2004-2005 reported having done so for better salary or benefits. For those who left teaching in 2004-2005, nearly 15 percent cited salary related reasons (NCES, 2007).

Teacher wages have increased dramatically over the last forty years. Nevertheless, since the 1970s, they have fallen behind salaries in nonteaching jobs for individuals with similar qualifications. Lawyers, doctors, scientists, and engineers
Teacher Quality:  
Broadening and Deepening the Debate

earn substantially more, as do managers and sales and financial service workers (Corcoran et al., 2004). The opportunity cost of becoming a teacher, in terms of salary forgone in alternative professions, is high. However, teachers may work fewer hours and fewer days and may receive more attractive benefits, at least partially compensating for this forgone income.

In 2003-2004, the average base salary of regular full-time teachers was $44,400 per annum. Public school teachers on average earned considerably more than their private school counterparts, the former making $44,500 on average and the latter $31,700.12 Regular full-time teachers in rural or small towns had, on average, lower base salaries than their counterparts in urban fringe or large towns and central cities (NCES, 2006).

Teachers’ salaries increase with years of experience and additional education. The average salary of beginning teachers in 2004-2005 was $31,753 per annum. There is considerable variation across states, with new teachers making up to $39,259 per annum in Connecticut and approximately $24,872 per annum in North Dakota. If we look at all teachers, and not just new teachers, we find considerable statewide variation too. Average teacher salaries are the highest in Connecticut at $57,760 per annum and the lowest in South Dakota at $34,039 (American Federation of Teachers, 2007). Much of this variation in salary mirrors variation in the wages of nonteaching college graduates, and thus the differences in dollars overstate the differences in the relative wages (and thus appeal of teaching) across regions.

Within a state, there are differences across counties, and within counties, between districts. The within-county differences, compared with differences across states, more closely reflect differences in relative wages and thus in the appeal of teaching relative to other occupational choices. Thus, salaries can affect not only whether an individual chooses to become a teacher, but also where they choose to teach. In Florida, for instance, teachers with a bachelor’s degree as their highest degree earned anywhere between $32,283 and $45,613 in 2005-2006, depending on where they taught in the state (Florida Department of Education, 2006). In Santa Clara County in California, teachers with similar educational qualifications were paid $66,652 per annum in Alum Rock Union Elementary school district during 2005-2006, but $80,041 per annum the same year in a neighboring district.13

A number of factors explain the variation in teacher salaries seen above. For example, districts with greater resources have more money to spend on teacher salaries. Alternatively, a district could have greater demand for teachers because of policy preferences for smaller class sizes or more skilled teachers; they may be willing to spend the money they have available on more teachers instead of potentially increasing the quality of their teachers by spending more on wages per teacher. Salaries could also be higher in one district than another because the region does not produce many teachers or because the job opportunities for college graduates are very good in other fields and thus the district has to pay more to attract equally skilled individuals into teaching.

12 The figure for public schools excludes charter schools.

13 California Department of Education: http://www.ed-data.k12.ca.us/ . The educational qualification referred to above is a bachelor’s degree and 60 Continuing Education units.
NONWAGE JOB CHARACTERISTICS

Salaries are only one criterion influencing individuals’ decisions about whether and where to teach. Non-wage job characteristics, including attributes of students, class size, school culture, facilities, teaching assignments, leadership and safety, also affect teachers’ choices and these characteristics often vary more dramatically across schools than do salaries.

Studies in Georgia, New York and Texas all find that teacher mobility is heavily influenced by characteristics of the student body, especially race and achievement (Scafidi, Stinebrickner, and Sjoquist, 2003; Boyd, Lankford, Loeb, and Wyckoff, 2005; Hanushek, Kain, and Rivkin, 2004). Georgia elementary teachers move from schools with higher proportions of minority students and from low-performing schools, but the latter appears to be explained by teacher preferences for fewer minority students. Texas and New York data, on the other hand, find that teachers prefer higher-achieving students even after controlling for student racial composition. Teachers, especially highly qualified teachers, are more likely to transfer or quit when teaching lower-achieving students. As further evidence of the weight some teachers put on student-body characteristics, when class size reduction in California increased the demand for teachers across the state, many teachers in schools with low-achieving students switched to schools with higher-achieving students (Shields et al., 2001).

While student characteristics are important by themselves, teachers also choose schools with more high-achieving and wealthy students because these schools often offer other characteristics that teachers prefer, such as better facilities or more preparation time. A recent survey of teachers in California, Wisconsin and New York found that schools serving large numbers of low-income students had a much higher incidence of inadequate facilities relative to other schools, evidence of vermin (cockroaches, mice and rats) in school buildings; dirty, closed or inoperative student bathrooms; inadequate textbooks and science equipment; and higher personal expenditures by teachers to compensate for insufficient classroom materials and supplies (Carroll, Fulton, Abercrombie, and Yoon, 2004).

The 2004-2005 TFS asked teachers who moved across schools why they moved. Table 1 shows that approximately 38 percent of teachers reportedly moved to another school due to a better teaching assignment. Interview studies also reveal that new teachers resent teaching subjects they do not know, subjects requiring extensive class preparation, being split between two subjects or teaching very large classes. While there is little evidence that these factors by themselves explain high turnover rates, it is likely that they cause stress and dissatisfaction, thereby precipitating teachers’ transfers and resignations (Johnson, Berg, and Donaldson, 2005).

School leadership is another important factor in teachers’ decision-making. In the 2004-2005 TFS, more than 37 percent of teachers indicated that this was an important factor in their decision to switch schools. Similarly, for teachers who left teaching altogether, Ingersoll and Smith (2003) found that of the 29 percent of leaving teachers who cited dissatisfaction as their reason for leaving, more than three-fourths linked their quitting to low salaries. However, the next two most important factors were student discipline problems and lack of support from the school administration. Two recent studies in New York and North Carolina also
find that teachers’ assessment of their schools’ administration is one of the most important factors predicting the turnover of early career teachers (Boyd et al., 2008; Clotfelter et al., 2008).

### Table 1: Percentage of Public School Teacher Movers Who Rated Various Reasons as Very Important or Extremely Important in Their Decision to Move to Another School: 2004-2005

<table>
<thead>
<tr>
<th>Reason for Moving to Another School</th>
<th>Percentage of Teachers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opportunity for a better teaching assignment (subject area/ grade)</td>
<td>38.1</td>
</tr>
<tr>
<td>Dissatisfaction with support from administrators at previous school</td>
<td>37.2</td>
</tr>
<tr>
<td>Dissatisfaction with workplace conditions at previous school</td>
<td>32.7</td>
</tr>
<tr>
<td>Higher job security</td>
<td>19.1</td>
</tr>
<tr>
<td>Dissatisfaction with changes in job description or responsibilities</td>
<td>18.3</td>
</tr>
<tr>
<td>Dissatisfaction with opportunities for professional development in previous school</td>
<td>12.8</td>
</tr>
<tr>
<td>Did not have enough autonomy over classroom at previous school</td>
<td>10.4</td>
</tr>
</tbody>
</table>


Teacher peers also affect teachers’ decisions. In a study of California schools, Shields et al (2001) find that credentialed teachers complained of the lack of professionalism of those who were not credentialed and the resulting instructional burden they had to carry to compensate for the teaching inadequacies of their colleagues.

Differences across schools in non-wage attributes of the job will be particularly important when there is little variation in wage to compensate, as is the case in large urban districts in which all schools operate under the same salary schedule. Policies that attract effective administrators, increase preparation time, decrease class size, or provide funds to renovate facilities can improve working conditions and thus help to equalize the distribution of teachers across schools.
LOCATION

In addition to wages and working conditions, school location has a strong influence on the distribution of teachers. Most teachers appear to prefer to teach near where they grew up or in districts and schools that are similar to the ones they attended as students. Of all public school teachers who chose to move from one school to another between 2003-2004 and 2004-2005, 26.2 percent cited closeness to home as a very or extremely important factor in their decision to move. Of those who left teaching, 11.2 percent cited changing residence as very or extremely important.

Sixty-one percent of teachers who entered public school teaching in New York State between 1999 and 2002 started teaching in a school district located within 15 miles of the district where they went to high school. Eighty-five percent entered teaching within 40 miles of their high school. Even when teachers go far away to college, they tend to come home to teach (Loeb and Reininger, 2004). A recent study using the NELS 1988-2000 data set and Common Core of Data finds that these results are consistent nationwide: teachers are indeed local (Reininger, 2006). Further, compared with college graduates in nearly 40 other occupations, teachers were significantly more likely to live locally eight years after high school graduation (Reininger, 2006). Cannata (2007a) argues that teachers tend to sort themselves into schools that are socially proximal to them, in terms of race and class, and resemble the schools they attended as children. She finds that teacher candidates tend to have a clear notion of where they want to teach and where they do not, despite knowing little about these schools. Thus, she concludes, even though teacher candidates espouse preferences for specific school characteristics, such as beginning teacher support, the eventual decision on where to teach is based more on feelings of familiarity, comfort, and fit (Cannata, 2007b).

Teachers’ preferences to teach close to home or in similar settings pose serious concerns for urban districts since these tend to be net importers of teachers. Urban areas do not produce as high a proportion of college graduates as suburban areas. Using schools with large minority enrolments and large percentages of students receiving free and reduced-price lunch as proxies for difficult-to-staff schools, Reininger (2006) finds these schools produce significantly lower percentages of students earning bachelor’s degrees—a prerequisite for teaching. As a result, schools in these regions need to attract teachers from other regions, for which they have to pay a premium to get equally qualified candidates. If they are unable to find qualified candidates, then they will be forced to hire from a less-qualified pool of applicants.

ENTRANCE REQUIREMENTS

In addition to factors affecting the appeal of a particular job, such as wages, working conditions and location, requirements for entry into teaching can also affect who goes into teaching and the distribution of teachers across schools. While teacher preparation and certification requirements could improve student outcomes by increasing skills and knowledge, they also impose costs on current teachers and would-be teachers for tuition and the opportunity cost of time. On the one hand, the willingness to incur such costs might signal those who are likely to be more motivated to teach; on the other hand, the costs per se could be prohibitively high for some, decreasing the potential pool of talented applicants. Licensure exams play a role similar to certification. While they have the merit of establishing a floor on the measured knowledge teachers must have, if the tests are unable to effectively
distinguish between better and worse candidates, or assess applicants on material unrelated to student learning, they may exclude teachers who might have been very effective in the classroom (Boyd, Goldhaber, Lankford, Wyckoff, 2006).

Until recently many schools, particularly those serving high concentrations of students in poverty, staffed their classrooms with uncertified teachers, despite the fact that in theory certification was required of all teachers. As an example, in New York City in 2000, 35 percent of teachers in the highest-poverty quartile of schools had failed the general knowledge certification exam the first time they took it and approximately half of all new teachers held a temporary license (were not certified to teach). As described above, NCLB changed the landscape, requiring that all students be taught by a “highly qualified” teacher by the end of the 2006-2007 school year. At least partially as a result, between 2000 and 2005 there was a remarkable narrowing in the gap in teacher qualifications between high-poverty schools and low-poverty schools in New York City. By 2005, only some 10 percent of new teachers in the highest-poverty quartile had failed their certification exam on the first attempt (Boyd et al., 2007a).

TEACHER HIRING PRACTICES

Factors that affect teachers’ decisions are only one side of the story. Factors affecting demand for teachers are important as well. Teacher hiring practices, for example, explain part of why some schools and districts end up with better teachers than others. A recent study by the New Teacher Project in three large urban districts in the Southwest, Mid-west and Eastern regions and one mid-size urban district in the Midwest found that some schools that appeared difficult-to-staff did not have a problem attracting teachers, but they did have a problem when it came to actually hiring them. Although there were between 5 to 20 times as many applicants as available positions in these districts, with up to 37 percent of the applicants in difficult-to-staff subjects such as math, science, special education and English Language Learners, each of the districts failed to make offers until mid to late summer. By that time, many of the applicants (31-60 percent) had withdrawn their applications. Of those who had withdrawn, 50-70 percent cited late timelines as a major reason for taking another job. Furthermore, the study indicates that applicants who withdrew from the process were significantly better qualified than new hires in terms of the likelihood of having a higher undergraduate GPA, a degree in their teaching field and completed educational coursework (Levin and Quinn, 2003). This suggests that districts with effective hiring practices such as aggressive recruitment strategies and spring job offers are likely to end up with higher quality teaching staff even if initially faced with the same pool of applicants. These districts are able to recruit their top choices while other districts are left with teachers who could not find jobs elsewhere.

Principals also do not always have the information needed to accurately assess teacher quality and judge future performance. In a recent paper, Jacob and Lefgren (2006) argue that while principals are able to identify the best and the worst teachers in their schools, they are not able to identify where the rest fall in the ability distribution. Principals, according to this study, also discriminated systematically against male and untenured faculty.

Liu and Johnson (2006) stress the importance of “information-rich” and timely hiring processes in improving the match between teachers, schools and teaching
assignments. In a survey of new teachers in California, Florida, Massachusetts and Michigan, they found that the hiring process relied heavily on reviews of paper credentials and interviews. Importantly, schools and districts rarely observed a candidate’s teaching. In much the same way, applicants rarely got much experience of the school they had applied to. Although most new teachers met with the school principal during the hiring process, very few interviewed with current teachers or met with students to get a feel of the school culture and requirements. As a result, new teachers in these states formed only a moderately accurate picture of what their job likely entailed, increasing the chances for job-related disappointments and turnover.

The timing of the hiring process might be the most severe impediment to information-rich hiring processes. Many new teachers are hired in summer, when school is not in session, teachers are unlikely to be available for interviews and classes cannot be observed in action. Further, Liu and Johnson (2006) found that approximately a third of new teachers in California and Florida were hired only after the school year had started, when principals were in a rush to fill a position, teachers were busy with their classes and there was little time for an informative hiring process. The combination of these factors underscores the difficulty—and necessity—of achieving effective hiring practices.

**BUREAUCRATIC HURDLES**

The problem of suboptimal staffing is driven, at least in part, by bureaucratic and contractual requirements (Levin, Mulhern, and Schunck, 2005). Three district-level policies may be particularly important: vacancy notification requirements, teachers’ union transfer requirements, and late budget timetables and poor forecasting. Vacancy notification requirements allow resigning or retiring teachers to provide very late notice of when they intend to leave. In the study of hiring practices in four districts conducted by the New Teacher Project, three had a summer notification deadline or none at all, while one had a mid-May deadline. Late notification deadlines make it very difficult to know which posts will be available in September, which is when the school year typically starts. Local laws and union contracts make it possible for experienced teachers to ask for last-minute transfers. Further, many principals delay advertising vacancies for fear of being required to hire a transferring teacher they do not want. Finally, late state budget deadlines lead to chronic budgetary uncertainties as a result of which administrators do not know which positions will be funded in their schools (See Levin and Quinn, 2003; Jacob, 2007).

**COLLECTIVE BARGAINING AGREEMENTS**

Collective bargaining agreements also influence hiring and retention practices and may affect the distribution of teachers across schools. Rules in these contracts, for instance, often make it very difficult to fire tenured teachers even when they are performing poorly. To the extent that parents can exert power to have such teachers removed from their children’s schools, less-effective teachers may be more likely to end up in schools serving students with the fewest available resources and the greatest needs. Similarly, the least effective teachers may end up in poorly-performing schools if the administrators are less effective as well. The collective bargaining process may also distort the allocation of resources toward easily measured factors such as salary, with other important aspects of schooling such as working conditions, bearing the brunt. Since non-wage factors such as working conditions
are important in determining whether high-quality teachers will come to teach in low-achieving and poor schools, this over-emphasis on pecuniary measures may be detrimental. Finally, policies tend to standardize across schools—salaries are just one example. If the needs of some schools are much greater than that of others, such standardization might put high-needs schools at a relative disadvantage.

In summary, the differences in teachers across schools are systematic and often striking. A variety of factors combine to create these differences. On the supply side, wages, working conditions, location, and entry requirements all contribute to the variation. On the demand side, hiring practices, bureaucratic hurdles, and collective bargaining practices are all important factors.

**TEACHER MOBILITY AND TURNOVER**

Once the decision to teach is made, the next question is where to teach. Differences in the characteristics of teachers across schools get determined, to a large extent, by teachers’ initial choice of posting. From an aggregate nation-wide perspective, the magnitude of teacher turnover is not very large. Between 2003-2004 and 2004-2005, for instance, 83.5 percent of teachers stayed in the same school, while only 8.1 percent transferred between schools and 8.4 percent left teaching (NCES, 2007).14 From the perspective of the individual school, however, attrition can be substantial because attrition rates vary across schools.

<table>
<thead>
<tr>
<th>Year</th>
<th>Stayers (%)</th>
<th>Movers (%)</th>
<th>Leavers (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991-92</td>
<td>87.6</td>
<td>7.3</td>
<td>5.1</td>
</tr>
<tr>
<td>1994-95</td>
<td>86.3</td>
<td>7.2</td>
<td>6.6</td>
</tr>
<tr>
<td>2000-01</td>
<td>84.9</td>
<td>7.7</td>
<td>7.4</td>
</tr>
<tr>
<td>2004-05</td>
<td>83.5</td>
<td>8.1</td>
<td>8.4</td>
</tr>
</tbody>
</table>

Source: Compiled from NCES (2007). Teacher Attrition and Mobility: Results from the 2004-05 Teacher Follow-up Survey.

*Stayers are teachers who were teaching in the same school in the current school year as in the year before (base year). Movers are teachers who were still teaching in the current school year but had moved to a different school after the base year. Leavers are teachers who left the teaching profession after the base year. Note, this does not rule out the possibility of their re-entering teaching at a later date.

14 New teachers are more likely to leave than more experienced ones. While this might be because teaching turns out to be somewhat more difficult than expected, it is important to note that data on recent college graduates show that young workers tend to switch jobs more, regardless of occupation.
From Table 2 we see that while the percentage of those who move to another school (henceforth called movers) has been fairly stable over the years, the percentage of those who stay on in a school (henceforth called stayers) has been decreasing gradually. Those who leave the teaching profession altogether (henceforth called leavers) has been rising steadily.

**CHARACTERISTICS OF MOVERS AND LEAVERS**

Younger teachers tend to leave a given school or the teaching profession more frequently than older ones. Between 2003-2004 and 2004-2005, for instance, 14.7 percent of teachers under age 30 years had moved to another school, while 9 percent had left teaching altogether. For teachers between age 40 and 49, on the other hand, only 7.1 percent had moved to another school and 5.3 percent had left teaching altogether.

According to TFS data, between 2003-2004 and 2004-2005, the category of teachers with no full-time teaching experience was the most likely to move out of a school as well as leave teaching altogether. For teachers with full-time teaching experience, those with 1-3 years of experience were both the most likely to move to another school as well as leave teaching altogether. The corresponding figures for teachers with more experience are lower; for instance, for teachers with 10-19 years of experience, 6.3 percent moved to another school, while 5.5 percent left teaching.

The difference by sex is not striking, although a larger percentage of female teachers left teaching altogether between 2003-2004 and 2004-2005 than male teachers. White teachers relative to black and Hispanic teachers had the lowest percentage of movers and leavers for the same period. The data suggest that Hispanic teachers had the highest percentage of movers, while black teachers had the highest percentage of leavers.

Not surprisingly, teachers whose base salary was $30,000 per year or less were the most likely to move to another school or leave teaching compared with teachers who earned more. With regard to main teaching assignment, special education teachers were the most likely to switch schools as well as leave teaching between 2003-2004 and 2004-2005. This stands in contrast to the period between 1999-2000 and 2000-2001 when special education teachers were among those least likely to leave teaching (though not among those less likely to move to another school).

Teachers who have a regular or standard certification type are the least likely to move to another school or leave teaching altogether. 7.2 percent of teachers who had a regular or standard certification type had switched schools between 2003-2004 and 2004-2005, while 8.2 percent of them had left. Those with a provisional or temporary certification type were the most likely to move, while those who had none of the common types of certification were the most likely to leave.¹⁵

¹⁵ A probationary certificate is issued after an individual completes all the regular certification requirements except the completion of the probationary period. A provisional certificate is issued to individuals who are still participating in what states call “alternative certification programs”. Temporary certification requires some additional college coursework, student teaching and/or passage of a test before regular certification can be awarded.
Better-qualified teachers (but not necessarily more effective teachers) are also more likely to leave teaching, at least in some regions. In New York City, for example, there are considerable differences between teachers who stay on in a particular school and those who transfer or quit (Loeb and Reininger, 2004). Those who stay on in a particular school have failed the certification exams twice as often as those transferring to another district (Boyd, Lankford, Loeb and Wyckoff, 2005). Moreover, the latter are twice as likely to have attended a highly competitive college, and half as likely to have attended a less competitive college. New York City teachers who quit teaching in New York State are also substantially more qualified than those who remain in terms of their test scores. For example, 20 percent of new teachers in the top quartile on the general-knowledge certification exam left high-achieving schools after one year, while 34 percent of those in low-achieving schools left after one year. By contrast, 14 percent of bottom-quartile teachers left high achieving schools after one year, and 17 percent left low-achieving schools.

More qualified teachers are also substantially more likely to leave schools having the lowest-achieving students. For example, of the new teachers hired in New York City’s lowest-achieving schools in 1996–1998, 28 percent scored in the lowest quartile on the general-knowledge certification exam. Of those remaining in the same schools five years later, 44 percent had scores in the lowest quartile. In contrast, 22 percent of the new teachers in the higher-achieving schools were in the lowest quartile, which increased to only 24 percent for those remaining after five years.

Teacher mobility also varies by geographical region and community type. Turnover rates in the Northeast region of the country are lower than in other regions, and larger schools face fewer turnovers than do smaller schools. Urban areas tend to have a slightly higher turnover rate than suburban areas in general, but there are certain urban areas where the situation is particularly bad. For example, in New York City approximately 62 percent of teachers switch schools within five years compared with 54 percent in the suburbs. Thirty-five percent of New York City teachers leave teaching altogether within five years compared with 25 percent of teachers in the suburbs.

Turnover rates in schools with higher proportions of African-American and Hispanic students are higher than in schools that are predominantly white. Scafidi, Stinebrickner, and Sjoquist (2003) find that Georgia elementary teachers move from schools with higher proportions of minority students and from low-performing schools, and that the latter is explained by teacher preferences for fewer minority students. Hanushek, Kain, and Rivkin (2004), using a similar model and Texas data, find that teachers prefer higher-achieving students even after controlling for student racial composition.

**IMPLICATIONS OF TURNOVER**

Teacher turnover may affect student learning in several ways. First, in high-turnover schools, students may be more likely to have inexperienced teachers who we know are less effective, on average (Rockoff, 2004; Rivkin, Hanushek, and Kain 2005; Kane, Rockoff, and Staiger, 2006). Second, high turnover creates instability in schools making it more difficult to have coherent instruction. This instability may be particularly problematic when schools are trying to implement reforms,
as the new teachers coming in each year are likely to repeat mistakes rather than improve upon implementation of reform. Third, high turnover can be costly in that it takes time and effort to continuously recruit teachers.

Transfer and quit behavior would be especially worrying if more effective teachers had higher attrition rates. This does not appear to be the case. Using data on a large urban school district in Texas, Hanushek and Rivkin (2006) find no conclusive evidence suggesting that more effective teachers, in terms of student test score gains, have higher exit rates. They find that those who exit are in fact less effective, on average, than non-movers, both in that district and in general. Further, those who move between schools within the same school district are, on average, less effective than those who do not. They go beyond average performances and compare the quality distributions of teachers who either change schools or exit public schools to get a more nuanced picture of what is happening with teachers at the top and bottom end of this distribution. They find that the distribution of these teachers falls distinctly below the distribution of those who stay, indicating that at every level, it is the less effective teachers who are more likely to change schools or exit public schools.

Their finding is echoed in a recent study of new teachers in New York City schools (Boyd et al., 2007a). This study also found no reason to believe that those who exited were better than those who stayed. Specifically, they found that first-year teachers identified as being less effective in improving student test scores had higher attrition rates than those identified as more effective. They found that it was relatively ineffective teachers, on average, who transferred within New York City; again, however, averages mask important variation. For teachers transferring from a given low-performing school, the more effective ones tended to transfer to schools with fewer low-scoring and nonwhite students, exacerbating the inequities in teacher quality across schools.

**POLICY INFLUENCES ON THE TEACHER WORKFORCE**

In the 2003-2004 school year, 74 percent of all public schools had teaching vacancies. Of the schools with vacancies, 16.4 percent reported having to hire a less than fully qualified teacher (NCES, 2006). Vacancies were highest in special education (67.4 percent), followed by English language arts (57.1 percent) and then math (55.6 percent). In each, the shortages were most pronounced at the secondary level and in urban schools (versus suburban and rural schools). While only 8.1 percent of schools with vacancies in ELA found it very difficult or were unable to staff their schools, the numbers for special education and mathematics were much higher at 29.2 percent and 28.8 percent, respectively. This section looks at the impact of different supply-side and demand-side strategies that aim to improve teacher labor market outcomes.

---

16 They measure teacher quality by looking at value-added in terms of standardized average student test score gains.

17 A final point on transfer and quit behavior. Exit decisions could just as well be driven by an especially unruly class in a particular year or a personal emergency. In fact, approximately 21 percent of teachers who quit teaching between 2003-04 and 2004-05 cited family or personal reasons as being very important in their decision to quit (NCES 2007).
INCENTIVES

As discussed previously, teachers’ salaries are important in the decision to teach and the decision to stay in a particular school. Nearly 17 percent of teachers who moved from their base school between 2003-04 and 2004-05 reported better salary and benefits as being very or extremely important in their decision to change schools. Approximately 14 percent of those who left teaching in the same period cited salaries and benefits as being at least very important (NCES, 2007).

Teachers’ salaries can be increased in two ways: (a) across-the-board increases in salaries, and (b) targeted increases, for example, by focusing on difficult-to-staff schools and difficult-to-staff fields. The economic argument for increasing the pay of all teachers already content to work in a given school is weak. Since it is unlikely that such schools will face staffing difficulties, it makes sense to target resources at teachers in difficult-to-staff schools and difficult-to-staff subject areas. It might also be beneficial to target higher salaries to more-effective teachers.

Many states and a large number of school districts are pursuing pay-related methods to recruit and retain highly qualified teachers. While retention bonuses are the most widely used of these methods, a few states offer housing incentives and a few offer signing bonuses to new teachers. Most of these policies are, however, not targeted at increasing the quality of the teaching force in shortage fields or in high-poverty or low-performing schools. Of the 35 states providing retention bonuses for teachers in 2003, only five targeted teachers in high-need schools (Loeb and Miller, 2007).

The evidence on the effectiveness of pay-related incentives on retaining teachers and improving student performance is small and mixed. The Massachusetts Signing Bonus Program for New Teachers, which started in 1998, combined a national recruitment campaign, $20,000 in signing bonuses and a seven-week “fast-track” certification program, but met with limited success in its stated goals. Twenty percent of the first cohort of bonus recipients left teaching after one year, and more than 50 percent of its second cohort ended up not teaching where policymakers said they should—in 13 state-designated, high-need school districts (Fowler, 2001). In 2001, North Carolina began giving $1,800 in annual bonuses to teachers in specific fields (math, science and special education) for middle or high schools serving low-income or low-performing students. This program mildly increased the retention of teachers, but it also suffered from complicated eligibility requirements and implementation problems (Jacob, 2007).

Incentives can also directly target success, rewarding teachers or schools that seem most effective. In a study of Dallas’ school-based accountability program, where every member of the staff of the most effective schools was rewarded, Clotfelter and Ladd (1996) found that the pass rates of students in the city increased relative to five other large Texas cities. Figlio and Kenny (2006), using data from the National Education Longitudinal Survey and their own survey conducted in 2000, however, found that test scores were higher in schools that offered individual-level financial incentives but not in schools that offered indiscriminate merit pay. While they were able to demonstrate that students learned more in schools in which individual teachers received financial incentives as reward for superior performance, data limitations prevented them from making causal linkages from their findings.
Incentives can also take the form of reduced costs of entry into teaching. Teachers have traditionally entered teaching after taking courses in four broad areas—foundational courses, pedagogical courses, subject-matter knowledge courses and field experiences—during either their undergraduate education or their master’s program. Many states, in an attempt to reduce the cost of entry for college graduates interested in teaching, now allow them to take alternative route programs with fewer course requirements prior to beginning teaching. Forty-seven states and the District of Columbia have some form of alternative-route program to recruit, train and certify teachers (Boyd et al., 2007b). Many states rely heavily on alternative routes for teachers. New Jersey, Texas and California, for instance, obtain more than one-third of their new teachers from alternative routes (Wyckoff, 2006).

Alternative route programs typically allow teachers to enter the classroom by delaying or bypassing many of the requirements for entry that are part of traditional teacher preparation programs. These programs require teachers to be college graduates and approximately 80 percent of them require demonstration of subject matter knowledge by completing coursework, passing an exam or some combination of the two. This apart, they vary greatly in requirements. ITeachTexas, a statewide alternative certification program in Texas, for instance, is a web-based alternative certification program which does not require any onsite pre-service meetings. The New York City Teaching Fellows Program (NYCTF), on the other hand, requires an intensive onsite seven-week pre-service training session.

The most commonly studied alternative route program, Teach for America (TFA), is better able to recruit teachers with stronger qualifications than those recruited through the traditional route. For instance, in 2003, TFA had 16,000 applicants, most from highly selective undergraduates, for 1,800 available slots. As a result, the program could be highly selective in terms of teacher qualifications; this is not true of all alternative route programs. Studies of the effectiveness of TFA teachers have found they are equally effective, or more effective, than other teachers in math, although the results for reading are less positive. For instance, in a randomized evaluation of the program in 17 schools in Chicago, Los Angeles, Houston, New Orleans, and the Mississippi Delta, researchers from Mathematica found that, although the average TFA-led student increased his or her rank in math by 3 percentile points over the course of a year, the average non-TFA student registered no change. In contrast, there was no difference between the average TFA and non-TFA student in reading gains, with both having registered an increase of 1 percentile. TFA teachers in the sample differed from non-TFA teachers considerably in terms of selectivity of college, education-specific training, certification and experience (Decker, Mayer, and Glazerman, 2004).

Similar, though not quite as positive, results hold for the New York City Teaching Fellows (NYCTF). Early estimates suggest that Teaching Fellows are less effective in their first year of teaching but that the differences in student achievement between NYCTF teachers and traditional teachers diminish with experience (Boyd, Grossman, Lankford, Loeb, and Wyckoff, 2006). Both TFA and NYCTF include substantial recruiting efforts as well as efforts to continuously improve, which makes it difficult to generalize the findings to alternative route programs, many of which may be less selective and put less effort into quality. In a recent study of alternative certification programs, Humphrey and Wechsler find a great deal of variation both between and within alternative certification programs, leading them to question
the worth of comparing different alternative certification programs. Further, the individuals who take up these programs have considerably different backgrounds, school placements and learning outcomes, making comparisons across programs problematic.

Districts have also been trying out various strategies to recruit people into teaching, especially minorities and people who belong to difficult-to-staff neighborhoods. Typically these involve partnerships between K-12 school districts and local colleges to encourage students to enter teaching or scholarship and loan forgiveness programs for candidates who commit to teaching for a certain period (Jacob, 2007). Broward County Public Schools in Florida, one of the five largest school districts in the United States initiated the Urban Teacher Academy Project (UTAP) to address a major challenge that faced the district: the need for 13,000 new teachers over the next ten years. The program recruits students when they are as young as 14 years old, grooms them in teaching techniques, classroom theory and pairs them with teacher mentors. After high school, they move on to community colleges and universities for a four-year, tuition-free teaching degree with a guaranteed job at the end. The program not only generates a larger number of teachers, but by drawing students from difficult-to-staff schools and minority areas, it also creates a teaching force that is unlikely to face culture shocks when it goes back to those schools to teach.\(^{18}\) Evidence on the success of these programs, in terms of student achievement and teacher quality and retention, remains sparse. A recent review of research by analysts at RAND and the Education Commission of the States found very little research on the impact of recruitment strategies employed in most states and districts (Jacob, 2007).

**REGULATING ENTRY**

Incentives are not the only way to influence the teaching workforce. One of the most common tools policymakers use to regulate the teaching profession is certification requirements. Most teachers in the United States are certified. For instance, in 1999-2000, 94.4 percent of public elementary and secondary teachers were certified in their main teaching assignment. In theory, certification keeps individuals who are likely to be poor teachers out of the classroom. The evidence on the effect of certification is, however, mixed. Recent studies in New York City and North Carolina found that students of certified teachers learned more, on average, than did students of uncertified teachers, though a similar study in Florida found no difference (Boyd et al., 2006a; Goldhaber, 2006; Harris and Sass, 2006). Similarly, studies in New York and North Carolina found that teachers who passed their certification exam (the Liberal Art and Science Test in New York and the Praxis II in North Carolina) showed higher student achievement in math. For example, teachers who passed the Praxis II produce, on average, student achievement gains ranging from 3 to 6 percent of a standard deviation higher (in math) than those who failed (Goldhaber, 2007). Comparing the effect of this gain to that produced by experience, the study found that the average teacher who failed the test, were he/she allowed to teach regardless, would likely produce the same level of math achievement in his/her second or third year of teaching as a novice teacher who passed the test. (Goldhaber, 2007). The study also shows how test cut-off criteria

can generate a number of false negatives (individuals who fail to pass the test but might have been high-quality teachers) and false positives (individuals who make the cut-off might turn out to be poor teachers), calling into question the signal value of certification tests (Goldhaber, 2007). Raising cut-off scores might also be detrimental if it reduces the supply and racial/ethnic diversity of the prospective teacher pool (Gitomer, Latham, and Ziomek, 1999; Angrist and Guryan, 2004).

**SUPPORTING TEACHERS**

In addition to regulating teachers, policy makers and educational leaders can affect the teacher workforce through policies that support teachers’ development. Surveys have found that the lack of support services rank high in teachers’ decisions to quit teaching (Jacob, 2007). Of teachers who changed schools between 1999-2000 and 2000-2001, 33.4 percent of new teachers (one to three years of experience) reported dissatisfaction with support from administrators as being very important in their decision to move. Almost 21 percent of those who left teaching during the same period reported dissatisfaction with job description or responsibilities as a very important factor in their decision to quit. Nearly 15 percent said that a very important reason for quitting was related to not feeling prepared to implement or not agreeing with new reform measures. Many districts have, as a consequence, adopted programs aimed at providing support, guidance and orientation services to elementary and secondary teachers as they begin their teaching career. The goal of these programs is to reduce teacher attrition by making teaching more manageable (Smith and Ingersoll, 2003).

Participation in induction programs increased during the nineties. In 1990-1991, approximately 40 percent of new teachers had participated in a teacher induction program; by 1999-2000, 80 percent had participated in an induction program (Smith and Ingersoll, 2003). Induction programs typically involve meetings, informal classes for new teachers and the formation of new-teacher peer support groups. Mentoring programs typically pair new teachers with experienced ones, although the details vary across programs. In a review of ten studies on induction and mentoring programs, Ingersoll and Kralik (2004) find empirical support for the claim that induction programs for new teachers and, in particular, mentoring programs have a positive impact on teachers’ decision to stay in the same school and continue in the teaching profession. Using 1999-2000 SASS data, Smith and Ingersoll (2004) find that certain types of activities, such as having a mentor from the same field and having common planning time with other teachers on instruction, were more effective in reducing turnover than other types of activities such as the provision of seminars or classes for beginning teachers (Smith and Ingersoll, 2004). A study of 141 teachers in New Mexico who participated in a teacher mentoring program found that the attrition rate was only 4 percent annually compared with the statewide average rate of 9 percent. In an analysis of the Beginning Teacher Support and Assessment Program (BTSA), a mentorship program in California, Vilar, and Strong (2005) found that in addition to reducing teacher attrition rates, the program resulted in aggregate reading scores for students of new teachers being comparable to those of mid-career teachers.

Professional development programs provide teachers with continuing education opportunities once they have joined the profession. These programs encompass traditional workshops, in-services, graduate coursework, school-based teacher
study groups, mentoring relationships, and advanced credentials such as that provided by the National Board for Professional Teaching Standards (NBPTS). Of teachers who moved from one school to another between 2003-2004 and 2004-2005, nearly 13 percent reported dissatisfaction with opportunities for professional development opportunities in their previous school as a very important reason for their decision.

Unfortunately, the research literature does not provide a clear understanding of the extent to which professional development programs improve student achievement. Although specific professional development programs have shown positive effects on student learning in randomized trials, there is considerable variation in the quality of professional development programs, and, on average, professional development programs do not appear to benefit students (Hill, 2007). Teacher self-reports of the quality of their own professional development experiences are not encouraging. In a recent study, only 20 percent of science teachers and 25 percent of math teachers said that their professional development program had changed their teaching practices (Horizon, 2002).

**SELECTIVE RETENTION**

According to an informal survey of the human resources departments in several large urban districts, less than 1 percent of the teaching workforce is dismissed each year (Jacob, 2007). Yet, selective dismissal or, similarly selective promotion, could affect the teacher workforce. In a recent study using data from New York City schools, Gordon, Kane, and Staiger (2006) argue that it is possible to predict the performance of a teacher in later years from student achievement scores in the first two years of teaching. On average, a teacher whose students make above average gains is likely to produce such gains in later years; similarly, a teacher who performs badly in the first two years is unlikely to improve dramatically. Making somewhat conservative assumptions about the costs of replacing ineffective teachers, they conclude that denying tenure to the bottom quarter of new teachers would substantially improve student achievement. This study, however, does not account for the potential change in teaching such a policy might facilitate, including a need to compensate teachers for the additional risk and the potential for undesirable narrowing or targeting of instruction (Jacob, 2007).

**LOOKING BEYOND TRADITIONAL PUBLIC SCHOOLS**

Charter schools and private schools may offer insights into teachers’ preferences and how to develop policies to attract and retain effective teachers. In a case study of 40 charter schools in Arizona which had completed their fifth year of operation in 1999-2000, Gifford and Ogle (2000) found that in general charter schools aimed to hire staff that had a philosophical connection with the school. They also sought less experienced teachers with the expectation that it would be easier to train and assimilate new teachers into the school environments.

The literature more generally identifies three reasons teachers are attracted to charter schools. First, teachers perceive charter schools as offering increased freedom, flexibility, and empowerment (Ascher, Jacobowitz, McBride, and Wamba, 2000; Finn et al., 2000; Koppich et al., 1998; Wohlstetter and Griffin, 1998). Second, teachers want to work in schools that share similar educational philosophies
(Finn et al., 2000; Hill et al. 2001; Koppich et al., 1998; Wohlstetter and Griffin, 1998; Vanourek et al., 1997). Finally, teachers believe charter schools offer smaller classes (Finn et al., 2000; Vanourek et al., 1997). Nevertheless, Podgursky and Ballou (1997) found teacher turnover to be significantly higher in charter schools relative to public schools. A study conducted by NEA concluded that charter school teachers were dissatisfied with their salaries and the lack of job security (Koppich et al., 1998).

Private schools share characteristics with charters. Researchers have found that religious beliefs and moral training at Catholic schools contribute to a sense of community and common purpose that improves teacher efficacy and morale (Bryk and Lee, 1993). This said, the percentage of private-school teachers leaving teaching since 1988-1989 has remained consistently higher than the share of public-school teachers leaving teaching. This higher attrition may be driven either by teachers decisions to leave (e.g., because of relatively low wages) or by greater administrative flexibility to dismiss teachers. Ballou (1996) suggests that private schools are more successful in retaining the best of their new teachers because of greater flexibility in structuring pay, more supervision and mentoring of new teachers, and freedom to dismiss teachers for poor performance. This may well be true, but there is no research to date that verifies or contradicts this proposal.

Schools outside of the United States can also provide useful insights because of the great variation in approaches seen throughout the world. In a review of the research on teacher labor markets in developed countries, Ladd (2007), for instance, finds that in most developed countries teachers’ relative pay is higher than in the United States. That being said, she finds no clear relationship across countries between teacher salaries and student achievement.

CAVEATS

It is important to emphasize that the findings from many of the studies cannot be treated as definitive, but only suggestive. Some of them fail to establish causality because they are unable to estimate the counterfactual, that is, what would happen in the absence of the particular intervention being studied. This happens for several reasons. First, it is not always clear what the counterfactual means. For instance, does it mean being taught by the average teacher in the district or by the least effective teacher? Second, even if one can establish the effect of a particular teacher on a group of students, it is not always possible to extrapolate those findings to an entirely different group of students (Murnane and Steele, 2007). It is also hard to disentangle contextual effects (such as school and classroom effects) from teacher effects in many studies. Because teachers choose where they hope to teach, it is likely that teacher assignment is related to student, classroom and school characteristics. This makes it difficult to distinguish statistically between effects that are due to teachers per se, and those that are due to characteristics of the students’ classroom, school, and district environment. Large, longitudinal data sets that follow students over time and match them to their schools and teachers have substantially increased our ability to sort among possible causes for the relationships that we see.
Promising Lines of Future Research

Three overlapping sets of categories are useful for framing promising lines of future research on teacher labor markets. First, the differentiation of (a) supply, (b) demand, and (c) institutions and contexts in which supply and demand jointly determine the workforce provides a simple framework for considering teacher labor market research. For example, research on voluntary attrition of teachers focuses on the supply side, while research on which teachers a district chooses from a pool of applicants focuses on the demand side. Studies of teacher contracts fall at the interaction of supply and demand. Often researchers are unable to distinguish supply and demand factors. For example, studies of the distribution of teachers across schools often cannot sort between the preferences of teachers and those of hiring authorities.

A second categorization of research focuses on the career stages of teachers: (a) pre-teaching, (b) early career, (c) middle career, and (d) late career. The pre-teaching period includes the recruitment, selection, and pre-service preparation of teachers. The early-career period includes mentoring and induction, monitoring and evaluation, retention (both general and strategic) and effectiveness. The area of retention is quite broad given that it can include factors such as working conditions that affect teachers’ choices of whether to stay, as well as the decisions of school officials whether or not to renew contracts. The middle-career period includes many of the same factors as the early career period but has a number of important differences. As an example, because most of these teachers are covered by due-process guarantees, schools and districts need different approaches for strategic retention. Professional development for these teachers also is likely different than that for novice teachers. In addition, the opportunities for job differentiation and advancement can affect the career decisions of mid-career teachers. The late-teaching period includes among other factors the retirement decisions of teachers.

A third categorization of research that can be useful distinguishes the type of research. For example, we could differentiate: (a) describing labor market dynamics, (b) developing and substantiating theories about the mechanisms driving the trends and relationships observed, (c) developing instruments for measurement, and (d) evaluating programs. Each of these areas of research can contribute to our knowledge of teacher labor markets but each provide different types of information and require different expertise to implement.

In almost every joint category, whether it is descriptive studies of the demand side factors affecting the retention of early career teachers or policy evaluations of changes in retirement benefits on teachers (supply-side) decisions to stay in teaching, there is plenty of room for new research. The following discussion focuses on four areas, one for each stage of the teaching career, that may be particularly productive to explore.

The first area is evaluation of the causal effect of recruitment and costs (particularly in time) of pre-service preparation requirements on teacher supply. Currently, Teach for America and the New Teacher Project have fundamentally changed the pool of individuals interested in becoming teachers. These programs have both reduced the time requirements for entry and put substantial effort into recruitment and selection. Traditional teacher preparation programs put very little emphasis
Teacher Labor Markets and Teacher Labor Market Research
• Susanna Loeb and Tara Béteille

on recruitment. Differentiating the importance of these two factors would inform policy development, especially in combination with research that identified selection criteria and pre-service preparation experiences that improve teachers’ effectiveness in the classroom. Unfortunately, it is rarely easy to assess the causal effects of policies. Our inability to do so is in part a consequence of how policies and practices are implemented; they are rarely implemented in a manner that allows for rigorous impact evaluation. We are in luck when cut-offs allow for regression discontinuity designs, or random assignment occurs through policy choices, lotteries or staged implementation. However, this is rare and often does not coincide with the most pressing questions. Finding ways to convincingly estimate causal models is a research challenge.

For early career teachers, one useful line of research would illuminate the demand side of teacher evaluation and strategic retention or dismissal. At this point, we lack even descriptive data on the extent to which schools and districts counsel-out or dismiss ineffective teachers. Because selection before entry appears, at best, to limit the numbers of very ineffective teachers but does not distinguish well above the left-hand tail of the distribution of teacher effectiveness, it is important to identify early-career teachers who are likely to be ineffective in the long-run and to somehow encourage them to leave. This is particularly imperative because of the greater difficulty of dismissing teachers once they are covered by due-process guarantees. Recent changes in how principals renew contracts of early-career teachers suggest that changes in routines can substantially affect the resulting teaching force in schools; for example making nonrenewal the default instead of renewal. Although evaluation of these types of changes would be useful, providing a description of the current state of affairs is at least as important. Some recent research shows that less-effective teachers, on average, are more likely to leave teaching in the early years, even though the number of official dismissals is low. We do not know the extent to which this differential attrition is driven by the demand side or the supply side.

For mid-career teachers, one useful area of research would develop better instruments for measuring teacher effectiveness. Currently researchers rely heavily on student test score gains. Often these tests are only available annually for grades three through eight in math and reading. Developing measures that capture other student gains for these teachers, student gains for other teachers, or other effectiveness measures such as student, peer or school leader evaluations or validated measures of teaching practice is essential for improved research on teacher effectiveness. With such measures we could, for example, better evaluate the effectiveness of professional development approaches and better describe the extent to which teachers who move into positions of leadership or who leave schools altogether were more or less effective in the classroom.

For late-career teachers, there is very little research to date on retirement decisions. This research is beginning, largely led by Michael Podgursky at the University of Missouri, but our knowledge is still in the early stages. To date we do not even have good information on policy variation. What are the retirement benefit policies across districts and states? We need to know these basic facts to evaluate the effects of these policies. Retirement plans are not the only policies for which we lack information. An important reason why it has been difficult to discern the impact of many policy interventions is that large micro-education data sets gather
very little information on the policy variables we are interested in, and it is costly for individual researchers to collect this information directly.

In summary, there is great opportunity to contribute to our understanding of teacher labor markets. Although the research is large and expanding, there are holes in our knowledge that meaningfully limit practitioners’ and policymakers’ ability to draw on evidence when making decisions that affect schools and students.

REFERENCES


Malloy, C. L., & Wohlstetter, P. (2003). Working conditions in charter schools:


NCES (2007). Teacher Attrition and Mobility: Results from the 2004-05 Teacher Follow-up Survey.


Many children spend more of their waking hours in classrooms than they do at home; within these settings they are exposed to experiences that for better or worse, intended and unintended, shape their development—they may learn to read, write, and think critically; they make friends and have to face the inevitable challenges of peer relationships; and they are oriented increasingly to become productive, independent members of a larger society. Interest is keen in the extent of these classroom effects, the methods of producing and reproducing them at various levels of scale, and understanding the mechanisms responsible for them. In fact, it could be argued that these interests in classroom effects are among the core focus of educational policy, practice, and research at the present time. The scale is staggering, with millions of classroom teachers working each day in public education settings, more than 200,000 of whom are new entrants to the profession each year, and with recent reports indicating 87 percent of whom leave the profession before a decade of experience (Pianta and Allen, 2008; Loeb and Béteille, 2008; Anderson, 2008).

The focus of this chapter is the consequences for policymaking of research on teachers’ classroom interactions with students, particularly standardized observations of those interactions. We address implications not only for education policy that aims to improve student achievement (such as No Child Left Behind), but also research policy. Our overall goal is to advance classroom interactions and processes as a legitimate focus of education policy (Tseng and Seidman, 2007; Shinn and Yoshikawa, 2008), based on what we know from research using standardized observations of those processes.

To preview, we argue that advances in theory, in measurement, and in intervention have led to the possibility that policy can reach into classrooms more directly than it has heretofore, taking shape, for example, in metrics for a “highly effective teacher” that rely neither on the proxies of degrees or experience that bear only indirectly (Gordon, Kane, and Staiger, 2008) and often not at all (Pianta and Allen, 2008) on student outcomes nor on the tautology that effective teachers are those who produce achievement gains (Rivkin, Hanushek, and Kain, 2005; Rockoff,
which runs the risk of too narrowly defining an important social asset. Rather, we argue that it is now feasible to experiment with policy that focuses on direct assessments of teachers’ performance in the classroom as an instructor, socializer, motivator, and mentor while demonstrating explicit evidence that these assessments account for student learning gains.

We also argue that research could advance the impact of classrooms on student performance if standardized observations of teachers’ classroom practices were more routinely and strategically a focus or embedded in large-scale value-added studies.

Finally, we contend that placing observational assessment of teachers’ classroom instruction and interactions more squarely into the realm of policy and accountability frameworks would in turn spur research and development efforts in teacher training likely to lead to more regular and efficient production of effective teaching by training programs and through ongoing professional development supports that actually work. In the end, our interest is in how hiring, promotion, or reward structure in schools; certification and licensure at the state level; pre-service preparation of teachers; and funding of education science, if centered on performance-based observations of classroom teaching behaviors, could function in a coordinated fashion to more systematically and directly produce better teaching in classrooms.

We position this chapter as a direct response to years of research on education inputs and outputs that has been of two forms, which we argue are both lacking in their capacity to inform policies leading to improvement of classrooms: one form identifies connections between distal proxies for inputs and student test score outputs through large-scale econometric modeling of the “education production function,” but is uninformative as to the proximal mechanisms of production; and the second form produces rich descriptions of classroom processes using in-depth studies of cases, but has little bearing at much higher levels of scale (see Ladd, 2008). The models and methodologies for conceptualizing, measuring, and improving classrooms that rely on standardized, scalable, valid observational assessments of teacher performance, we argue, provide tools that complement econometric and case methods and hold promise for improving the effectiveness of American education.

**CLASSROOMS ARE THE LOCUS OF EDUCATION EFFECTS**

The last decade has witnessed a marked increase in research examining the complex nature of children’s experiences in classrooms and the ways in which these experiences uniquely contribute to children’s social, cognitive, and academic development. This increase in research has been driven by several factors. First, educational reform in general, accountability frameworks at the state level, and the No Child Left Behind Act in particular, have placed individual schools and teachers in the spotlight: requiring evidence of their role in producing student achievement. Thus, the attention of policymakers and school administrators has turned to the (admittedly too-thin) research base on links between classroom experiences and student outcomes.

This need for information on what happens in classrooms is exacerbated further as the results of widespread accountability and state standards-testing become used for identifying schools that do not meet established proficiency levels and rates of
performance. Witness the concern in any local paper or in state government when schools are not making “Adequate Yearly Progress” or fail to make accreditation standards. And reauthorization of NCLB is launching a range of discussions that shift the dynamic of accountability to focus on how inputs produce achievement through debate about effective or qualified teachers (e.g., Appalachian Regional Advisory Committee, 2005; Foundation for Child Development, 2005; Gordon et al., 2008; Hamre, Pianta, Mashburn, and Downer, 2007; Ladd, 2008; Perie, Moran, and Lutkus, 2005). Classrooms are the hotspot as states and districts scramble to find the right mix of curriculum, professional development, and instructional supports that will raise students’ achievement (e.g., Appalachian Regional Advisory Committee, 2005; Educational Testing Service, 2004; Lasley, Siedentop, and Yinger, 2006).

A second reason for focusing on classrooms is illustrated in the widespread and persistent arguments being made for state-supported pre-kindergarten as a means of addressing a range of social and economic concerns (Bogard and Takanishi, 2005). This evidentiary stream clearly demonstrates (more so than most effects-focused research) that classroom processes are the locus of early education program effects for high-risk young children. A host of studies describing results of randomized controlled trials and highly controlled quasi-experimental work, have shown quite clearly that significant gains in achievement accrue as a function of enrollment in pre-kindergarten (pre-k) classrooms, and that children’s interactions with teachers are in large part the source of these effects (see Mashburn et al., 2008).

We focus here in pre-k not as an end in itself, but illustratively; to the extent there is a body of well-controlled studies of educational effects of classrooms processes, it exists for pre-kindergarten educational programs. More widely, when schooling is viewed as a solution to address student performance gaps, social or economic inequality, or international competitiveness (Carroll, Fulton, Abercrombie, and Yoon, 2004), the rhetoric, evidence, and programs are focused on what accrues to students as a function of their engagement in classroom processes.

Not surprisingly, the third body of evidence in support of classrooms as a focus of policy draws directly from studies that indicate quite clearly that classrooms, and teachers, matter. In studies of large-scale statewide testing programs in which multilevel analysis has been used to isolate sources of variance to which achievement growth can be attributed (e.g., Nye, Kostanpoulos, and Hedges, 2004), classrooms are more often than not the greatest source of variation in what students learn and gain as a function of attending school (at least in achievement-related domains; it is not as clear with regard to social outcomes). Recent work, largely motivated by NCLB’s focus on highly qualified teachers and an assortment of concerns related to teacher licensing systems, has focused on teacher characteristics such as degrees, experience, qualifications, and test performance, and shown significant returns to achievement as a function of each of these features—and policy-relevant effect sizes when considering their cumulative benefits (see Ladd, 2008; Loeb and Béteille, 2008).

Classroom processes are implicated as significant moderators of treatment effects in highly controlled experimental work (Battistich, Watson, Solomon, Lewis, and Schaps, 1999). Within classrooms, differences in teachers’ implementation of treatments (e.g., curricula) appear to be the single biggest factor determining
effects on child outcomes (Domitrovich and Greenberg, 2004), even when teachers are instructed to deliver the intervention in a standardized manner and are provided with regular and intensive supports to maintain fidelity. Thus, in both value-added research and experimental studies not only is a good part of the value of attending school conveyed at the level of classrooms, but it also appears that teachers play a major role in determining the value of the classroom environment for student learning and development.

Thus, as the focus of research seeking to locate and maximize the benefits of education shifts to recognize and consider processes within classrooms that could account for such benefits, whether in value-added or experimental frameworks, the evidence shows that experiences with teachers can matter. Relatedly, new investments of research dollars and a targeted effort to engage researchers from other disciplines in the educational sciences have spurred interest in pre–k-12 classrooms, particularly among developmental psychologists. Thus, we now see the bridging of developmental and education science (Pianta, 2007) in ways that have produced marked gains in understanding classroom processes related to instruction, teacher-child interaction, and peer relations.

One thread of this connection can be traced to economists, who have been instrumental in advancing the science of value-added modeling of achievement gains, using longitudinal data to isolate effects of classroom-level experience for which, it is argued, causal influence can be attributed (see Ladd, 2008; Murnane and Steel, 2007). A second thread leading to interest in classrooms is evident when basic science disciplines, such as developmental psychology, move from lab to setting; for example, scientists interested in peer relations have moved from experiments in the lab to careful study of peer relations in classroom and school settings (Gifford-Smith and Brownell, 2003). In the case of peer relations, such movement to classroom settings has led to an understanding of how teachers set norms for aggressive behavior that have consequences for peer victimization and learning (Henry, Guerra, Huesmann, Tolan, VanAcker, and Eron, 2000) and how teachers’ use of language and responses to social cues of students of varying gender or ethnicity have consequences for motivation, learning, and social relationships in the classroom (Weinstein, 2008).

For the remainder of this chapter, we focus on one source of classroom effects on achievement—teacher-student interactions—and the consequences when an evidence base is accrued regarding the measurement, production, and improvement of interactions that are effective in producing achievement gains.

One reason for this focus is the evidence regarding teachers as a primary locus for schooling effects. The other reason is that for policymaking, the production of teachers (and presumably teaching quality) is of real concern, is under considerable regulatory control at the state and federal levels, and has extraordinarily high stakes attached to its success or failure—as just one example, see Pianta et al. (in press), who reports that in the thousands of U.S. classrooms taught by certified teachers, only 25 percent provide a level of instructional or emotional support consistent with the production of learning gains. As another, refer to the report by Clotfelter, Ladd and Vigdor (2007) that certain profiles of teacher characteristics have negative effects on achievement for certain student groups.
We argue that policies that place direct assessment of actual teaching as a central feature of accountability frameworks and provisions for equity of educational opportunity are likely to accomplish several interlocking aims that in a coordinated fashion could result in substantial shifts in the nature and quality of instruction, socialization, mentoring, and tutelage that takes place in classrooms. First, these policies will result in better teaching in classrooms; second they will lead to improved student outcomes; and third they will drive a research and development enterprise that will soon result in a science of the production of teaching and teachers—something that simply does not now exist.

TEACHER-CHILD INTERACTIONS MATTER

A major question occupying the attention of education science and policy has been to determine the degree to which placement in a particular classroom accounts for variability in children’s academic performance. Numerous studies have indicated that: 1) a significant portion of variance in student learning is explained at the classroom level (see Nye et al., 2004); and 2) deflections in the trajectory of student learning across years can be attributed to their experiences in specific classrooms (see Ladd, 2008; Rivkin et al., 2007).

Although these studies have been important in laying a foundation for inquiry into classroom effects, they fail to articulate specific processes that may lead to student learning and positive social adjustment. The problems with this approach are highlighted in recent debates about teacher education and teacher education research. Hanushek’s (2002) definition of teacher quality—“Good teachers are ones who get large gains in student achievement for their classes; bad teachers are just the opposite” (p. 3)—as well as much of the research using the value-added paradigm (see Gordon et al., 2006, for overview) provide virtually no guidance to the development of a scientifically driven agenda aimed at evidenced-based ways to produce good teaching, either through improved in-service support or teacher education (Cochrane-Smith and Zeichner, 2005).

Of course, some have argued that value-added models could be used to identify effective and ineffective teachers, who could then be observed to detect the behaviors that differentiate them (Gordon et al., 2006). Although this is a promising pursuit, make no mistake that it is fraught with serious methodological problems, notably what Garmezy, three decades ago, termed the “retrospective fallacy” (Garmezy, 1977). In research on the prediction of psychopathology, Garmezy noted the trend to identify carefully matched individuals with and without a disorder and then compare those pairs on measures of potential interest. Garmezy was careful to note that this strategy is flawed because a) it will always find a difference on some indicator (even if there is no detectable conceptual reason for the difference), b) it is not possible to attribute cause to indicators on which there is a difference because of problems with selection bias, and c) these indicators always perform more poorly (if at all) as discriminators in predictive models involving the entire population. Current interest in value-added models as a way of discriminating successful and unsuccessful teachers for further study may suffer from the same fallacy in terms of being a path for identifying factors causal to teachers’ putatively effective behaviors (or student achievement).
We argue that, on its own, value-added work as a means of identifying and producing effective teachers is inherently limited, both conceptually and methodologically because it relies on retrospective identification of effectiveness and has no conceptual framework for hypothesis testing and experimentation about putative mechanisms linking inputs to outcomes (Pianta et al., in press). The more important challenge is to identify (perhaps with value-added models as one source) a set of conceptually sensible and empirically valid teaching practices, and assign students randomly to teachers exhibiting those practices; or to build programs (interventions) for training teachers to exhibit those behaviors and test their efficacy on teacher and student outcomes.

Defining and Understanding Classroom Environments
Consistent with developmental theories emphasizing the role of proximal process on development (Bronfenbrenner and Morris, 1998), we take the view that a key ingredient of any classroom environment is interaction between adults and students. This view of classroom environments excludes a focus on some aspects of classrooms that have been the focus of research, such as the availability of materials or curricula, or the number of students in the setting. However, it provides a broad, holistic view of the classroom environment that includes all types of teacher-student interactions, those that are socioemotional, organizational, and instructional in nature. As such, our view of classroom environments is inclusive of research focused on more discrete aspects of classrooms such as quality or effective teaching, learning environments, and student-teacher and peer relationships.

In school, children’s experiences in classrooms constitute the majority of the proximal processes responsible for achievement and other outcomes. Although students do spend time in other places (e.g., lunchrooms, playgrounds) the bulk of their day is contained within classroom walls and experiences at the classroom level appear to be most closely associated with student outcomes (Nye et al., 2004). School-level effects, such as school climate and culture, are distal factors, the effects of which are largely mediated through or that moderate classroom process effects. For example, when schools have a positive climate, in which faculty, staff, and students feel safe and supported, teachers and students are better able to recreate this environment in the classroom than if the school climate is marked by hostility and discord—thus, the effects of school climate may be mediated by interactions in the classroom (Deemer, 2004).

Similarly, findings from intervention studies provide strong evidence of substantial variability of implementation between classrooms (Jones, Brown, and Aber, 2008) even when the entire school is being directed to adhere to a standard intervention protocol. Mashburn (Mashburn and Pianta, 2008) argues that structural features of schooling or classrooms that are typically the levers engaged by policymakers (e.g., teacher education) do not exert direct effects on child outcomes but instead should be treated conceptually and analytically as moderators of the effects of interactions in the classroom setting.

To help organize the diverse literatures that might inform the task of describing teacher-student interactions in classrooms, Hamre and Pianta (2007) presented the CLASS Framework. The CLASS Framework is a theoretically driven and empirically supported conceptualization of classroom interactions organized into three major domains—Emotional Supports, Classroom Organization, and Instructional
Using Standardized Observation to Conceptualize, Measure, and Improve Classroom Processes • Robert C. Pianta and Bridget K. Hamre

Supports. Within each domain are a set of more specific dimensions of classroom interactions that are presumed to be important to students’ academic and/or social development (see Figure 1). The CLASS framework starts with an understanding of the nature and regulators of developmental change at a given period (Hamre and Pianta, 2007; Pianta and Allen, 2008), then applies that understanding in a developmentally informed analysis of classroom settings and teacher-child interactions, maps that understanding back onto the rich and deep literature on classroom teaching and educational effectiveness, and then organizes this analysis within a framework that could inform measurement. McCaslin and colleagues (2006) have taken a similar approach in measuring classroom-setting effects on student motivation. The result is a multilevel conceptual and measurement model that can be subjected to empirical tests and evaluation.

The CLASS framework is one of several descriptions of classroom environments or quality teaching put forth in the educational and developmental literatures (e.g., Brophy, 1999; Brophy and Good, 1986; Eccles and Roeser, 1999; Gage, 1978; Pressley, Roehrig et al., 2003; Soar and Soar, 1979). For example, Brophy (1999) describes 12 principles of effective teaching, including supportive classroom climates, opportunities to learn, curricular alignment, thoughtful discourse, scaffolding engagement, and achievement expectations, each of which are based on research findings and theories of teaching and learning. Others organize teachers’ practices into larger domains of teaching and classroom environments. Pressley and colleagues (2003) draw from their studies of effective teachers (e.g., Bogner, Raphael, and Pressley, 2002; Pressley, Allington, Wharton-McDonald, Block, and Morrow, 2001; Pressley, Roehig et al., 2001; Wharton-McDonald, Pressley, and Hampston, 1998) to suggest that effective teaching strategies can be organized into decisions regarding motivational atmosphere, classroom management, and curriculum and instruction. Similarly, Eccles and Roeser (1999) suggest that schooling is characterized by organizational, social, and instructional processes that help regulate children’s and adolescents’ development across cognitive, social-emotional, and behavioral domains.

The CLASS framework draws heavily from this theoretical and empirical work. There are, however, several distinctions between the CLASS Framework and other frameworks for studying classroom, one being that Hamre and Pianta (2007) propose a latent structure for organizing teaching behaviors that in most approaches are simply culled and categorized by type. The latent structure in the CLASS model poses explicit, testable hypotheses regarding the organization of meaningful patterns of behavior(s) that are tied to underlying developmental processes. This conceptualization is explained in Pianta and Allen (2008) as a way of thinking about classroom supports that begin with an understanding of the developmental salience of various inputs and behaviors. The CLASS model then starts with a developmental analysis of settings, which is different than thinking of classrooms as only serving educationally framed aims (e.g., achievement) and then seeking to define what is good teaching. Importantly for policy, one implication of the CLASS conceptualization of classrooms is that the desirable outcomes related to classroom experience are not limited to performance on achievement tests, but rather involve a range of competencies that could be valued or assessed, and which could mediate or moderate achievement gains.
In this structure, meaningful units of teacher-child interaction are organized by patterns, which in turn are the basis for identifiable and scalable dimensions of interaction. These dimensions are then organized into one of three broad domains of classroom supports. For example, the domain of emotional supports includes three dimensions: positive classroom climate, teacher sensitivity, and regard for student perspectives. Organizational support includes effective behavior management, productivity, and instructional learning formats. And the instructional support domain includes concept development, quality of feedback, and language modeling dimensions. The organization of these dimensions into these three broad domains has been tested and validated for grades pre-k to fifth and a somewhat different set of dimensions, within the same hypothesized three-domain structure, have been specified for grades 6-12. The dimensions included in the current pre-k through fifth-grade framework have received the most empirical support but are not exhaustive; there are likely be other dimensions that could fit within each domain, and as we suggest, dimensions may vary with developmental period or epoch.

Within each of these dimensions are posited a set of behavioral indicators reflective of that dimension, which in turn are operationalized at various anchor points on a 1-7 scale using specific behaviors and interaction patterns that can be reliably observed in a specified window of time. For example, positive classroom climate includes observable behavioral indicators such as the frequency and quality of teacher affective communications with students (smiles, positive verbal feedback) as well as the degree to which students appear to enjoy spending time with one another. This detailed, multilevel conceptualization of teacher-child interactions in classroom environments, moving back and forth between broad theoretically-based domains to very specific behavioral indicators, is a way of organizing and linking, under one framework, approaches to classroom assessment and observation that are driven by different methodological and theoretical considerations and heretofore have remained distinct. It also presents a set of testable hypotheses concerning the organization of behavior at varying levels of abstraction.

A key distinction between the CLASS framework and others is that there is empirical support for this organization of classroom processes that derives from large-scale studies of actual classrooms (Hamre et al., 2007). It is not trivial to emphasize that these features of the supporting evidence—that it is empirical in nature and drawn from many thousands of observations—enhance the extent to which this evidence can have some bearing on policy and make this evidence base somewhat unique. Drawing from a sample of over just under 4,000 preschool to fifth-grade classrooms that were a part of several large, national and regional studies, Hamre and colleagues (2007) first examined observational instruments and sorted observed dimensions into the domains described by the CLASS framework. They then used confirmatory factor analysis to examine the extent to which this organization of classroom interactions was consistent with actual observations in these settings and to test alternative organizational structures. Results suggested adequate fit of the three-factor model and that the fit of this model was superior to a one- or two-factor model. These findings provide evidence that the three-domain structure suggested by the CLASS framework fits the natural variation in classrooms.

A final distinction between the CLASS framework and other conceptualizations of classrooms is that it was developed to apply to teacher-student interactions in
classroom contexts *across* grades and *across* content areas, from preschool to high school, and thus differs from work that has focused on more discrete groupings of grades, or has a specific focus on teachers’ content knowledge or delivery of content, when the focus is itself on content (the contribution to achievement gains of content-neutral CLASS ratings, content-specific behaviors, or teacher content knowledge remain open empirical questions). There are three reasons why the CLASS framework has not included, in a direct manner, content in the descriptors or scales it applies to classroom interactions. First, the system was conceptualized within a developmental psychology framework that posits that the key engines of development are the properties embedded in interactions between children and the setting, and that properties of these regulatory systems (such as feedback or reciprocity) are what drive development (Sameroff, 1983; Roeser, Eccles, and Sameroff, 1998). Second, CLASS was designed to be applicable across as many classroom settings as possible and to have utility as an observational tool even when the content focus of the lesson or instruction changed in unpredictable ways. Third, as we have had to address questions related to content, we have worked with individuals to identify and specify what would be desired in a content-focused scale (say, one focused on math in fourth grade), a large proportion of the behavioral indicators are pedagogical or interactive in nature and not content-specific (e.g., presents lesson in an organized manner). We refer readers to Hamre and Pianta (2007) for a more detailed description of, and rationale for, the domains and dimensions represented in this system of assessment.

We remind readers that CLASS encodes teacher-child interactions in terms of global ratings on dimensions represented in terms of seven-point rating scales, with such judgments typically being made after 20 minutes of observation. There are an assortment of questions pertaining to CLASS and other standardized observational protocols in classrooms about decomposing variance attributable to factors such as raters, time of day, time of year, length of observational window, and unit of analysis, and others (Raudenbush and Sadoff, 2008). Each of these parameters is treated as error in models attempting to isolate effects of teachers, and so to the extent they can be modeled in analysis of teacher effects, the precision of prediction improves as does effect sizes of teacher-related variables (Raudenbush and Sadoff, 2008).

In a series of analyses that cannot be fully documented here, but are available on request, CLASS ratings from multiple studies across multiple grades were examined in relation to these basic generalizable, theory-informed questions. The evidence suggests that indeed rater-related variance, even after raters pass stringent tests for interrater agreement and reliability, are substantial in these global dimensions of classroom interaction, but that such effects are attenuated with increased numbers of raters within a given data collection system, and pale in comparison to time of day and window effects for frequency-related codes (Chomat-Mooney et al., 2008). Indeed the evidence suggests that CLASS does capture aspects of teacher-child interaction that are stable across a specific day, across days, across students, and across content area of instruction, thus providing a reasonable estimate of features of a teachers’ behavior that appear stably characteristic of her interactions with students.
It is stunning, given the importance of classroom settings as vehicles for the transmission of knowledge and skill in our system of education, that little-to-no population level data exist pertaining to exposure of children and adolescents to particular classroom practices that are either known to relate to academic success or failure, desired on the basis of certain policies or values, or even hypothetically expected to relate to outcomes. Although, as we describe below, there is evidence emerging for early education and elementary classrooms, in secondary classrooms there is no current work which provides national-level, observational data on these environments. Two large national studies conducted over the past ten years do provide some of the first “epidemiological” data on preschool to fifth grade U.S. classrooms (Early et al., 2005; NICHD ECCRN, 2002, 2005; Pianta et al., 2005; Pianta, La Paro, and Hamre, 2008).

Overall these studies suggest that the average child is exposed to moderate levels of emotional support and classroom organization, and fairly low levels of instructional support throughout preschool to elementary school (Early et al., 2005; NICHD ECCRN, 2002, 2005; Pianta et al., 2005, 2008). In general, teachers are fairly positive in their interactions with students and examples of teacher or student negativity are relatively rare (NICHD ECCRN 2002, 2005). However, these interactions between teachers and students appear to be fairly impersonal, with very few instances in which individual students have positive, one-to-one interactions with their teachers.

For example, in fifth-grade classrooms, positive, individual interactions with a teacher occurred in only 1 percent of observed intervals across a school day (Pianta et al., 2008). The typical student also has few interactions with teachers around behavior management issues in schools, either positive or negative (NICHD ECCRN, 2005). However, one clear indication of problems in classroom organization comes from consistent findings that students spend a great deal of their time in classrooms without being exposed to any learning activity at all, ranging from 42 percent of the time in preschool classrooms to 30 percent of time in fifth-grade classrooms. Evidence on the quality of instructional supports is particularly concerning, with consistent evidence that children across grades are unlikely to be exposed to high-quality supports such as concept development and feedback. For example, by a ratio of nearly 5 to 1, children in fifth-grade classrooms are exposed to instructional activities (across any content area) that are basic-skill focused, in contrast to a focus on analysis, inference, or synthesis of information (Pianta et al., 2008).

Most notable in these and other studies, however, is the high degree of variability in classroom quality. A typical school day for some students includes spending the majority of time engaged in productive instructional activities with caring and responsive adults who consistently provide feedback and challenge students to think critically. For others, a typical day consists of spending most of the time sitting around, watching the teacher deal with behavioral problems, and engaged in boring and rote instructional activities such as completing worksheets and spelling tests (Early et al., 2005; NICHD ECCRN, 2002, 2005; Pianta et al., 2005, 2008).
These problems of inconsistent exposure to high quality classrooms are compounded by clear evidence of inequity. Students coming from disadvantaged backgrounds are more likely than their peers to be exposed to poor quality (Hamre and Pianta, 2005; Pianta et al., 2005). Further troubling is evidence that even the student lucky enough to experience a high quality classroom one year is very unlikely to be systematically exposed to high quality over a period of years, even if they remain in the same school (NICHD ECCRN, 2005; Pianta et al., 2008), suggesting that school-level resources such as professional development supports and school climate are insufficient to ensure high quality classroom environments. Taken together these studies suggest that very few of the students who are in greatest need of high quality classroom experiences receive them and the few that do are unlikely to receive them consistently, making it unlikely that the positive effects will be sustained.

Changes in Classroom Environments across Grades

Another area in which very little is known concerns patterns of exposure to various classroom supports as children move classrooms from year to year. Most research that does exist in this area has looked specifically at transition periods, either from preschool to kindergarten, into middle school, or into high school (Anderman and Midley, 1997; Ferguson and Fraser, 1998; Eccles, Flanagan, Lord, and Midgley, 1996; Hamre et al., 2007; Rimm-Kaufman and Pianta, 2000). For example, work on the transition to kindergarten has focused on ways in which classrooms become more oriented toward learning and less oriented toward social development (Hamre et al., 2007).

Others have provided evidence of the shift in goal orientations and student-teacher interactions from elementary to middle school, with students typically rating teachers as more distant and less supportive of autonomy in middle school—just at the time when, developmentally, young adolescents may be most in need of those positive supports from teachers (Anderman and Midgley, 1997; Ferguson and Fraser, 1998; Eccles et al., 1996). A recent study provides evidence that this shift is reported by teachers as well; as children move from kindergarten through sixth grade, there is a general pattern of decreased relational connections (both positive and negative), particularly around fifth grade (Jerome, Hamre, and Pianta, in press). Finally, during the transition from middle to high school, students report decreased levels of engagement with the content of schooling (Yair, 2000). However, we know little about normative shifts outside of these crucial transition periods.

In sum, despite the importance of teacher-student interactions in classrooms to student achievement and other outcomes, and that such processes can be reliably assessed and have been shown to predict to achievement, only one major national study provides information on the nature and quality of these key educational assets for large numbers of students. Moreover, this study, although large, is not representative of national or state demographics and under-represents students of color or from families of varying language or economic backgrounds. In comparison, there are literally dozens of large-scale epidemiological studies of health-related inputs.
**Improving Teacher-Student Interactions**

A final set of issues to be addressed concerns how and the extent to which student-teacher interactions can be improved. By and large these issues concern the nature and effectiveness of professional development and certification. At present, there is a consensus that the vast machinery of teacher licensing, certification, entry-level and advanced degrees, and in-service professional development play far too little a role in producing achievement gains for students (or effective teacher-student interactions as evinced in the Pianta et al. study described above). The one exception to this pattern is the set of findings showing that teachers with degrees and courses in math tend to produce higher math scores for their students (see Goldhaber, 2008). Rather, the general conclusion to be drawn from this work is that the metrics and processes used in the training and certification of teachers are very hard to link, either directly or indirectly, to student achievement gains or to observed effectiveness of teachers’ interactions in classrooms. One argument that has been advanced as an explanation for this is that the metrics used to regulate, certify, license, and produce teacher quality are not tied to direct observation (measurement) of what teachers do.

As accountability frameworks penetrate teacher education programs in a variety of ways—data systems that link teacher preparation experiences to value-added modeling of student achievement, studies of program ‘output’ and there is a noticeable push for teacher education itself to be more strongly grounded in evidence and research. By nearly every visible marker, the pressure is on teacher education to deliver supports and training that can demonstrably produce more effective classroom learning environments for students. Critically, this movement of attention toward teachers and teaching comes with a mandate for rigor, for careful and through evaluation, and for the capacity to discriminate knowledge that can generate verifiable solutions at state- and district-level scale from opinions and impressions. Yet despite the acknowledged importance of this need, we could not find a single empirical paper that one would think is essential to evidence-based decisions and policies in the production of effective teaching that was not based on student achievement and market forces.

In a recent controlled evaluation, we examined the extent to which a system of professional development supports that was organized around and focused on CLASS-based measurement and description of teacher-child interactions, using video exemplars of high-quality interaction as well as individualized consultation and coaching (MyTeachingPartner, 2004; Pianta et al., in press) was effective for producing changes in teacher-child interactions and child outcomes. The initial MTP trial tested whether it was possible to provide professional development experiences for early educators that enhanced their interactions with children.

Two models of training were implemented for two years and compared in this longitudinal study: 1) **Web Only**, in which teachers received workshop training in the curriculum and had access to the video exemplars, and 2) **Consultancy**, in which teachers received the same workshop training as the other condition but also received the regular, web-mediated consultation. Every two weeks, teachers videotaped their implementation of an instructional activity and shared this footage with the research team. In the consultancy condition, web-mediated interactions between a teacher and consultant then focused on: a) observing the video footage and identifying a teacher’s behaviors with students and their effects; b) problem-
Using Standardized Observation to Conceptualize, Measure, and Improve Classroom Processes • Robert C. Pianta and Bridget K. Hamre

solving to identify and implement alternative approaches as needed and receiving feedback on such attempts; and, c) establishing a nonjudgmental and nonevaluative supportive relationship with a knowledgeable individual (Hadden and Pianta, 2006).

We compared improvement in specific dimensions of teachers’ observed interactions across two intervention conditions (Pianta et al., in press), access to Web Only exemplars or Consultancy. Hierarchical Linear Modeling (HLM) estimated teachers’ growth trajectories for each of the nine CLASS dimensions of teacher-child interactions over the course of the school year. In terms of the main effects of intervention condition on improvement in the quality of teacher-child interactions, all associations were in the expected direction. Teachers exposed to Consultancy supports, in contrast to the Web Only, showed statistically significantly greater improvements in aspects of interaction that involved reading and responding to students’ cues (Teacher Sensitivity), using a variety of formats to actively engage children in instruction (Instructional Learning Formats), and intentionally stimulating language development (Language Modeling).

We also tested whether either of the intervention conditions was more effective in classrooms where a majority of children were from high poverty families. Results indicated that within the highest poverty classrooms, participating in the Consultancy condition was associated with positive changes in Teacher Sensitivity and Instructional Learning Formats while teachers participating in the Web Only condition declined over the year in the quality of their interactions on these dimensions. Thus, it appears that the level or intensity of supports a teacher might need to be successful in the classroom depends in part on the demand characteristics of the classroom itself.

Following evidence that the MTP intervention can alter teacher-child interactions, the next logical question is whether this change in classroom context has any echo in children’s social, academic, and language skills. We used two analytic approaches to address this question: experimental intent-to-treat and quasi-experimental treatment-on-the-treated. Taking full advantage of the RCT study design and two years of intervention implementation, we first evaluated preschoolers’ language and literacy outcomes across the three study conditions; these conditions include the Consultancy and Web Only groups described above, as well as a third control group in which teachers only received curriculum lesson plans (referenced to as Materials). Children’s assessed pre-kindergarten language and literacy skills were greater for children in classrooms in which teachers received individual consultation (Consultancy), with an effect size of $d = .21$, compared with the Web Only and Materials conditions (Fan, Pianta, Justice, Hamre, Downer, and Mashburn, 2008).

The quasi-experimental, treatment-on-the-treated analysis exclusively examined teachers in the Consultancy and Web Only conditions (teachers in the Materials condition did not have access to similar resources; Mashburn et al., 2008). During the intervention, these teachers varied in their use of three MTP web-based resources: language/literacy activities, a website that features video exemplars of high-quality classroom interactions, and expert consultation. We therefore examined the associations between teachers’ exposure to these resources and children’s development of language/literacy skills during pre-kindergarten. Controlling for relevant covariates (child, teacher, and classroom characteristics), children showed
significant gains in directly assessed receptive language skills when their teachers received Consultancy support compared with those who were in classrooms of teachers who received no such support. As expected, findings indicated a stronger effect of Consultancy support on children’s directly assessed print awareness and teacher-reported language/literacy skills in year two compared with year one.

In summary, there is evidence from our work and others’ (see Jones et al., 2006) that teacher-child interactions can be improved when interactions are the explicit focus of professional development, and that children make significantly greater achievement gains in language and literacy skills when their teacher receives support focused on their interactions with children. In short, basing professional development explicitly on a validated, observational assessment of teachers’ interactive behaviors in the classroom produces improvements in behavior but in children’s learning.

**IMPLICATIONS**

There is now a reasonable body of evidence, emanating from our work and that of others (see Gordon et al., 2008; McCaslin et al., 2006; Jones et al., 2006), that teachers’ performance in classrooms, in terms of their behavioral interactions with students, can be observed using standardized protocols, analyzed systematically with regard to various sources of error, and in turn prove to be valid for predicting student learning gains. They can also be changed (improved) as a function of specific and aligned supports provided to teachers, and that exposure to such supports is predictive of greater student learning gains. Although modest, these effects are robust and consistent across investigator groups, samples of teachers, and samples of students that vary by grade and socioeconomic and geographic background.

Below, we briefly summarize the implications of this work for policy.

We have argued that a major advantage of observational assessments of teachers for leveraging improvements in educational outcomes is that they can be directly related to the investigation and experimentation of specific interventions aimed at improving teaching. For this reason, these methods have considerable promise. Yet measurement challenges, some of which are noted above pertaining to psychometric issues, are not inconsequential. In addition to those challenges described earlier, observational assessments require technical supports that enhance efficiency and lower costs when used at scale. The questions related to psychometrics, efficiency, and costs compel attention and rigorous study, yet the investment in research related to assessments of such “inputs” pales in comparison to research investments in outcomes, specifically standardized tests. Nonetheless, recent Institute of Education Sciences requests for applications do include research on teacher effectiveness and specific topical areas on teacher quality; however, the assessment and measurement goals in the IES framework do not include research on assessment of teachers’ performance in classrooms.

It seems important, at least to us, that investment in measurement studies, cost-efficiency studies, investigations of the determinants and regulators of the quality of teacher-child interactions, and the value of teacher preparation programs for improving such interactions, could be key areas for research and development support. We can envision studies that would identify early predictors of teacher competence, effective supports that improve teaching, virtual reality environments
that accelerate teacher development, and networks of teacher preparation programs studying the natural history and course of teachers’ competence in these performance domains. Although this work could indeed be informed by value-added metrics of teacher quality and effectiveness, we suspect the route to eventual useful interventions and tools will be quicker and perhaps more efficient if focused on teacher-child interactions.

We recognize here the limitations of the work we have described above—measurement issues, small effect sizes, and logistic challenges that impede efficiency and scalability. Yet despite these challenges, the consistency and nature of the results suggests several implications for education policy.

In the realm of accountability, teacher quality and teacher effectiveness are critical “inputs” that counterbalance the focus on student achievement outcomes and potentially address value-laden issues such as equality of opportunity. Given the rather meager results related to teacher characteristics such as education, training, and experience as they relate to outcomes, the results presented herein with regard to observational metrics merit attention as a complementary feature of a comprehensive system.

With regard to production of effective teaching, using teacher preparation programs, credentialing and licensing systems, observational metrics have potential to provide anchors in actual performance that could be used to drive these systems toward higher levels of impact on student performance (providing the observation system is validated). Furthermore, if these systems used observations of performance as an outcome or competence metric, it is highly likely that collateral professional development supports would be developed that were oriented to producing performance at levels specified in licensing and credentialing systems. Finally, performance-based metrics anchored in observations in classrooms may also be compatible with market-based approaches to creating performance incentives, such as merit pay structures.

In sum, we argue that standardized observational approaches to measuring teacher performance represent a credible complement to the current focus on teacher credentials and degrees on the one hand, and value-added metrics of student performance, on the other. Furthermore, observational approaches may link more easily to systems of producing teaching that, in the long run and despite costs and logistic challenges, represent an alternative that has greater long-range benefits for building capacity and quality.
REFERENCES


Using Standardized Observation to Conceptualize, Measure, and Improve Classroom Processes • Robert C. Pianta and Bridget K. Hamre


Using Standardized Observation to Conceptualize, Measure, and Improve Classroom Processes • Robert C. Pianta and Bridget K. Hamre


Weinstein, R. S. (2008). Schools that actualize high expectations for all youth theory for setting change and setting creating. In M. Shinn and H. Yoshikawa (Eds.), *Toward positive youth development: Transforming schools and community programs* (pp. 81-101). New York: Oxford University Press.


Helen F. Ladd is the Edgar Thompson Professor of Public Policy Studies and professor of economics at Duke University. hladd@duke.edu

Susanna Loeb is a Professor of Education and Director of the Institute for Research on Education Policy and Practice at Stanford University.

Tara Béteille is a PhD Candidate in the Economics of Education program, School of Education, Stanford University.

Robert C. Pianta is the director of the Center for Advanced Study of Teaching and Learning, Curry School of Education, University of Virginia.

Bridget K. Hamre is a research scientist at the Center for Advanced Study of Teaching and Learning, Curry School of Education, University of Virginia.