Everyone's Doing It, But What Does Teacher Testing Tell Us About Teacher Effectiveness?

Dan Goldhaber*
University of Washington and the Urban Institute

ABSTRACT: This paper explores the relationship between teacher testing and teacher effectiveness using a unique dataset that links teachers to their individual students. My findings show a positive relationship between some teacher licensure tests and student achievement. But, they also suggest that states face significant tradeoffs when they require particular performance levels as a precondition to becoming a teacher: some teachers whom we might wish were not in the teacher workforce based on their contribution toward student achievement are eligible to teach based on their performance on these tests, while other individuals who would be effective teachers are ineligible.

* The University of Washington, CRPE, 2101 N. 34th Street, Suite 195, Seattle 98103.
Email: dgoldhab@u.washington.edu
Phone: 206-616-8793
Fax: 206-221-7402

Acknowledgements – This research is based primarily on confidential data from the North Carolina Education Research Center at Duke University, directed by Elizabeth Glennie and supported by the Spencer Foundation. The author wishes to acknowledge the North Carolina Department of Public Instruction for its role in collecting this information. The author gratefully acknowledges the Carnegie Corporation of New York, the Ewing Marion Kauffman Foundation, and an anonymous foundation for providing financial support for this project. The author also wishes to thank Jacob Vigdor, Michael Podgursky, and Ronald Ehrenberg for comments on an earlier version of this article; Carol Wallace and Michael DeArmond for editorial assistance; and Dan Player and Mike Hansen for excellent research assistance. The views expressed in this paper do not necessarily reflect those of the University of Washington or the study's sponsors, and responsibility for any errors rests solely with the author.

I. Introduction

Education research dating all the way back to the *Coleman Report* (Coleman, 1966) shows that of all the school-related factors that affect student achievement, teacher quality is the most important, and further that quality varies considerably among teachers.[1] Indeed, some studies suggest that having an excellent teacher instead of a poor one can translate into an additional year's worth of learning growth (Hanushek, 1992). As a result, there is great deal of interest in understanding teacher quality and the ways in which various education policies, for better or worse, affect it.

A primary screen used by all states in an attempt to guarantee a minimal level of teacher quality is the teacher licensure system (also commonly referred to as "teacher certification"): individuals who want to become public school teachers must meet certain requirements. All states, for example, require teachers to hold a bachelor's degree and have some training in pedagogy in order to be licensed. Most also require that teachers have training in the subject they teach and some kind of student teaching experience. Teachers typically also have to pass background checks and state-mandated tests before they can work in the classroom (Rotherham and Mead, 2004). This paper focuses on the most straightforward of these requirements, teacher testing - if you can't meet or exceed a 'cut score' on a licensure test, the state deems you ineligible to teach.

[1]Both Rivkin, Hanushek, and Kain (2005) and Rockoff (2004) report similar findings: they estimate that a 1 standard deviation increase in teacher quality is estimated to raise student achievement in reading and math by about 10 percent of a standard deviation. The magnitude of this effect is enormous relative to other educational resource interventions—being roughly equivalent to a reduction in class size of 10 to 13 students, depending on the grade level (Rivkin et al., 2005).

Despite the popularity of teacher testing as a policy (all but two states require teachers to pass some kind of licensure test[2]) there is a great deal of uncertainty about their value as either a screening tool to limit the number of low-quality teachers in the workforce or a signaling device to be used by local school systems for making hiring decisions. Both theoretical work (Stigler, 1971) and empirical work in other contexts (Kleiner, 2000) suggests that licensure, in general, is not a guarantee of service quality. There is relatively little empirical work linking teachers' scores on licensure tests to student achievement, and the magic pass/fail test line varies by state and is typically set by expert consensus panels, not empirical data. In the absence of good evidence about the relationship between these teacher tests and measures of teacher effectiveness, its not possible to judge the extent to which states' use of licensure tests allows ineffective teachers into the workforce or screens potentially effective teachers out of the workforce.

This paper explores the relationship between teacher testing and teacher effectiveness, as measured by their value-added contribution to student learning gains, using a unique dataset from North Carolina that links teachers to individual students in grades three through six over a 10-year period (1994-95 through 2003-04). These data allow me to account for the nonrandom distribution of teachers across schools and classrooms as well as the nonrandom attrition of teachers from the workforce, which is essential as these factors could bias estimates of the relationship between teacher testing and effectiveness. I also exploit a natural experiment that arises from time-series and cross-sectional variation in state cutoff scores in order to explore the

[2] As of 2006, Iowa and Montana are the two states that do not require tests for licensure. This information was compiled from the ETS website on individual state requirements (http://www.ets.org/portal/site/ets/menuitem.22f30af61d34e9c39a77b13bc3921509/?vgnextoid=d378197a484f4010VgnVCM10000022f95190RCRD) as well as department of public instruction websites for individual states. .

extent to which changes in the cutoff result in the exclusion of potentially highly effective teachers from the labor market or the inclusion of ineffective teachers in the market.

On balance, I find a small positive relationship between some teacher licensure tests and student achievement, however the point estimates are generally quite small in the specifications that account for the nonrandom sorting of teachers across students. This suggests that states face significant tradeoffs when using these tests as a screening device: despite the testing, many teachers whom we might wish were not in the teacher workforce based on their contribution toward student achievement are nevertheless eligible because they scored well on their test. Conversely, many individuals who would be effective teachers are ineligible due to their poor test performance. However, this does not necessarily mean that states should dispose of licensure tests, as they may provide an important signal about teacher quality that local hiring authorities could weigh against other teacher attributes when making hiring decisions.

The paper is arranged as follows. Section II provides more background on teacher testing. Sections III and IV describe the analytical approach and the data used for this study. The results are presented in Section V, and Section VI offers some concluding thoughts on the policy implications of the findings.

II. Teacher Testing: Background and Evidence

At the heart of any licensure test requirement is the *exclusion* of certain individuals from the pool of potential employees: if you don't pass the test, you aren't in the pool. States implement teacher testing requirements in order to exclude individuals who would have been teachers of unacceptably low quality. Whether or not that happens, however, will depend on the distribution of quality among the test takers as well as the 'cut score' used to keep individuals out of the pool. Licensure testing may also have other effects on the teacher workforce besides

this screening function; for example, testing increases the cost of labor market production (Friedman and Kuznets, 1945), and thereby discourages people from becoming teachers. Indeed, research by Hanushek and Pace (1995) and Angrist and Guryan (2004) suggest this may be the case in teaching. Finally, testing may provide a signal of employee quality that could influence public school hiring decisions. This paper, however, will focus solely on the screening and signal values of licensure testing.

States began testing teachers as a condition of employment in the 1960s, but since then, states have increasingly enacted formal teacher-testing policies. In 1985, 31 states required teachers to pass a test as a prerequisite for employment (Flippo, 2002), as of 2006, a total of 48 states and the District of Columbia required potential teachers to pass tests which cover basic skills, content knowledge, and/or professional knowledge (Title II Technical Assistance, 2003). The great majority of these states use one or more of the Educational Testing Service's (ETS) Praxis series of tests (Title II, 2003).

While states place differing levels of emphasis on testing teachers, they tend to use a uniform *approach* to determining the 'cut score' that candidates must achieve to become a teacher. Typically, states rely on a panel of education experts who attempt to relate the minimum levels of content and teaching knowledge required of beginning teachers to what is measured by the various licensure tests—the resulting 'cut score' is where they deem a minimally qualified teacher candidate should perform (National Research Council, 2001).[3] Given this method, it should come as no surprise that states tend to make different decisions about what the appropriate cutoffs should be, even when they use the same test. These cut scores may be

[3] States that use the ETS Praxis series of tests use a Angoff-type (1971) model to determine cutoff scores. In the Angoff model, panelists base pass rates on estimations of the proportions of each minimally qualified candidate who could answer the question (National Research Council, 2001).

internally valid with regards to measuring particular skills, but no state uses a cut score that reflects scientific evidence about a particular level of teacher effectiveness.

It is worth noting that ETS, the main developer and purveyor of teacher tests, emphasizes that its teacher licensure assessments are designed to "measure the knowledge and/or skills thought to be important for beginning practice," and nothing more (*Proper Use of the Praxis Series and Related Assessments*, 2006). Given this, some have argued that evaluating licensure tests as signals of teacher *quality*, as I do here, is a questionable enterprise at best, as licensure tests are valid only to the degree that they accurately reflect the knowledge and skills actually required of beginning teachers (Jaeger, 1999). While not dismissing the technical difficulties involved in examining the link between licensure tests and teacher quality, the question of whether or not licensure tests provide a valid signal of quality is clearly an important one. Measuring content knowledge and skills for their own sake is of little use, unless they are somehow related to job performance. And, as a practical matter, states ostensibly use licensure tests as *de facto* quality screens in the teacher labor market.

Despite the insistence that such tests should not be evaluated as signals of quality, there is a small literature that suggests performance on these tests actually do serve as a good signal of teacher effectiveness, though estimates of the strength of the teacher test-student achievement relationship vary significantly across studies.[4] Most likely, this inconsistency is due to the aggregation level of the data analyzed (Hanushek, Rivken, and Taylor, 1996), and the extent to which the research accounts for the potential nonrandom match of teachers to their students (Clotfelter, Ladd, and Vigdor, forthcoming). Two recent studies that use disaggregated data find

[4] See, for example, Ehrenberg and Brewer (1995), Ferguson (1991), Ferguson and Ladd (1996), and Strauss and Sawyer (1986), Summers and Wolfe (1975) as examples of studies relating teacher test performance to student outcomes.

a consistent, though small, relationship between teacher performance on licensure exams (the Praxis tests used in North Carolina) and student achievement. Clotfelter et al. (forthcoming) find that a 1 standard deviation increase in teacher test-score performance is predicted to increase 5th grade students' achievement by 1 to 2 percent of a standard deviation. Goldhaber (forthcoming), focusing on the elementary grades (3rd – 5th), finds that a 1 standard deviation change in teacher test-score performance is predicted to increase student test scores by about 1 to 4 percent. Neither of the above papers, however, explore whether licensure tests are differentially predictive of teacher quality at different points in the test distribution, or account for the possibility that sample selection or nonrandom attrition from the teacher labor market may bias their results.

III. Theoretical Framework and Analytic Approach

If licensure tests work as the screening devices they are intended to be (it is worth remembering that these tests constitute only one of several components of most states' licensure systems), they will exclude relatively few individuals who would have been highly effective teachers, and exclude most of those who would have fallen below an accepted threshold of effectiveness. But, it is inevitable that some individuals who score well enough to make it into the teacher labor market will end up being quite ineffective teachers; these individuals are referred to as *false positives*. At the same time, other individuals who would have made very effective teachers score poorly on the tests and are therefore excluded from teaching; they are referred to as *false negatives*.[5] The number of false positives and false negatives will depend on how closely licensure performance and teacher quality are correlated across the licensure test distribution: if the relationship between the two is strong, there will be relatively few false

[5] See Goldhaber (2004) for a more comprehensive discussion of this issue.

positives and false negatives, but, conversely, if the correlation is weak, there will be significantly more.

From a policy perspective we might be interested in two aspects of these tests: their efficacy as a screening device (as they are currently used by states), and the information they provide as a signal of teacher quality. To assess the efficacy of using these tests as a screening device, I estimate the following basic educational production function:

$$(1) \qquad A_t = \alpha A_{t-1} + \beta STUDENT + \delta PASS + \eta CLASS$$

The left hand side of the equation (A_t) is the achievement of student *i* in year *t*. The model includes controls for achievement in a prior year, A_{t-1}; a vector of individual student characteristics, STUDENT; an indicator for whether a teacher passed or failed a licensure exam based on a particular standard, PASS; and a vector of classroom variables, CLASS.[6] The main focus of interest is in the estimate of δ, which is identified by the comparison of teachers who pass the licensure test standard to those who do not, and therefore serves as a measure of the average differential in student achievement between teachers who pass and those who fail a licensure exam, holding constant the other variables in the model.

As described in Section II, the Educational Testing Service (ETS) claims that its commonly used teacher tests may not have great predictive power away from a given state's cut

[6] Analyses of the value-added of various school and teacher effects is generally based on one of three empirical specifications: one, like that specified in equation 1, where the dependent variable, a measure of achievement, is regressed against a set of controls that includes a measure of prior achievement; a second, where the dependent variable is the gain (the difference between a post test and some measure of prior achievement) in test scores regressed against a set of controls; and finally, a specification where achievement is regressed against a set of controls that includes student fixed-effects. In this paper I have experimented with all three specifications, and unless noted, the reported results for the teacher licensure test variables do not vary significantly from one specification to another (the magnitudes of the point estimates change slightly but the patterns of statistical significance do not).

score.[7] But there is no "national" cut score, and the fact that scores vary considerably between states (National Research Council, 2001) warrants a closer look at the predictive power of licensure tests along the entire performance distribution. This is certainly relevant for policy, given that a local school district, at the point of hire, may use a teacher's actual score (as opposed to just her pass or fail status) as an indicator of teacher quality.

To ascertain the signal value of teacher test scores, I estimate a variant of equation 1 that substitutes a vector of teacher tests, TEST, for the indicator for whether a teacher passes the states cutoff, and includes a vector of other teacher characteristics, TEACHER, since local districts, at the point of hire, have the benefit of assessing teacher test scores in the context of additional information about teachers (for instance, the type of certification they hold). In the following specification, the coefficient on TEACHER, λ, is identified based on variation in teacher test scores among teachers in the labor force and it reveals the signal value of the test:

$$(2) \qquad A_t = \alpha A_{t-1} + \beta STUDENT + \lambda TEST + \phi TEACHER + \eta CLASS$$

In carrying out the above analysis, there are at least three potential sources of bias. The first is that, with few exceptions (described in more detail below), teachers are only observed if they have met the minimum licensure standard. This creates a classic problem of sample selection that has been shown to lead to bias in other contexts, such as estimates of the impact of wages on labor supply (Killingsworth, 1983) or SAT performance on college grades (Vars and Bowen, 1998). This is problematic for identifying the screening value of licensure tests, as one

[7] The quote is: "the lack of an exact value for the highest score obtainable follows from the fact that the PraxisTM test scores are intended to be interpreted with reference to the passing score set by each state that uses the test in the process of licensing teachers. Because licensing decisions are, by law, meant to protect the public from harm rather than to allow selection of outstanding candidates, distinctions among test takers near the top of the score scale are not important for the use of the test in making licensing decisions. For more information, see the ETS website posted replies to questions:
http://www.ets.org/portal/site/ets/menuitem.2e37a093417f63e3aa77b13bc3921509/?vgnextoid=a2912d3631df4010VgnVCM10000022f95190RCRD&vgnextchannel=57ec253b164f4010VgnVCM10000022f95190RCRD.

might hypothesize that those teachers who make it into the labor market despite having not met the licensure test standard are likely to possess attributes valued in the labor market that are not measured in the dataset.

Fortunately for research purposes, states periodically change their required cut scores and grandfather teachers into the new standard. As a result, there are teachers in the data set, described below, who are in the workforce despite failing to meet a standard that was in place during their tenure. Furthermore, because states set different cut scores, I can explore the pass/fail signal for different points in the distribution that have been judged to be valid based on other states' standard-setting procedures. The sample selection issue does not lead to a bias in the signal value of teacher tests *among those teachers who are deemed eligible to teach* by virtue of having passed the test. However, including those who fail the test but are in the workforce nonetheless does have the potential to bias estimates of the signal value of the licensure tests, if the above argument about their value in the labor market is correct, since these individuals at the bottom of the test distribution are likely to have unobserved attributes that are positively correlated with student achievement.

Another potential source of bias is the nonrandom distribution of teachers across students. A significant amount of research (for example, Lankford, Loeb, and Wyckoff, 2002; Loeb, 2001) shows that more-advantaged students, in terms of family income and parental education, tend to be assigned to higher-quality teachers (as measured by such characteristics as experience, degree level, and test performance). Furthermore, this type of nonrandom matching is likely to produce upwardly biased teacher coefficient estimates (Clotfelter et al., forthcoming). To address this problem, I exploit the longitudinal nature of the dataset to estimate variants of equations 1 and 2 that include school or student fixed effects (the measure of prior achievement,

A_{t-1}, is omitted from the student fixed-effects models).

A third potential source of bias is that attrition from teaching is unlikely to be random. A large literature suggests that teachers with higher levels of demonstrated academic skills (measured by standardized tests) are far more likely to leave the profession (Hanushek and Pace, 1995; Murnane and Olsen, 1990; Podgursky, Monroe, and Watson, 2004; Stinebrickner, 2001; 2002). This can also be a source of bias if these academic skills are correlated with teacher quality. I address this issue by focusing on a sub-sample of novice teachers to see whether or not the estimates of the relationship between licensure test performance and teacher effectiveness appears to be different for this sub-sample.

IV. Data and Descriptive Statistics

I use data drawn from administrative records maintained by the North Carolina Education Research Data Center (NCERDC) for the North Carolina Department of Public Instruction of (NCDPI). These records include all teachers and students in the state over a 10-year period (covering school years 1994-1995 though 2003-04).[8] These data are unique in that they permit the statewide linkage of students and teachers (at the elementary level) and the tracking of both over time. They also include detailed student background information such as gender, race and ethnicity, parental education, disability and free or reduced-price lunch status, as well as performance on end-of-grade reading and math tests (described in more detail in North Carolina's *Standard Course of Study*- *cite?*), which are vertically aligned and explicitly designed to measure student achievement growth.[9]

The teacher data include degree and experience levels, licensure status, the college from which the teacher graduated, and the teacher's performance on one or more licensure exams.

[8] Student information for 4th and 5th graders for 1996-97 is missing from the dataset.
[9] The specific variables used in the various model specifications are reported in notes at the bottom of each table.

From the 1960s through the mid-1990s, individuals wishing to become teachers in North Carolina were required to pass the National Teachers Exams (NTEs). These exams were two hour multiple choice tests in specific specialty areas. The NTEs were replaced by the Praxis series of exams starting in the 1990s. The Praxis series include the Praxis I (reading, writing, and math)—which is considered to be a basic reading, writing, and mathematics skills test—and the Praxis II, which is a sequence of tests that focus on specific subject-matter knowledge and/or pedagogical preparation.[10]

As of July 1997, elementary school teachers in North Carolina (the focus of this study) were required to attain certain levels on specific Praxis II tests (0011 and 0012): the Praxis II Curriculum, Assessment, and Instruction ("Curriculum") test and the Praxis II Content Area Exercises ("Content") test; however teachers entering laterally into the workforce after this time could also meet the testing requirement with acceptable scores on either the NTE or the Graduate Record Exam (GRE). Some teachers in our sample also took the Praxis tests prior to the state's requiring it in 1997 (likely because they were required in another state), and a number of teacher records also include scores on the Praxis I because it has been required for admission into North Carolina-approved teacher preparation programs. As a result of the long-standing testing requirements in North Carolina, the data include information on some type of teacher test for over 91 percent of the North Carolina teacher workforce.

All teacher tests are normalized (to have a mean of zero and standard deviation of one) to place teachers on the same metric. The Praxis tests are normalized relative to all Praxis test takers in a given year based on national means and standard deviations for each Praxis test in

[10] Samples of Praxis tests may be downloaded from www.ets.org/praxis/download.html.

each year of test administration.[11] In theory, this normalization is not necessary as the tests are designed to be equivalent across years such that the timing of the test administration should not influence a candidate's performance (Understanding Your Praxis Scores 2005-06, 2005). Nevertheless, the normalization is useful in order to place teachers who took different Praxis tests or the NTE on the same metric.[12]

Some teachers have multiple test scores on their record, either because they took a single test multiple times (in order to pass a performance threshold), because they took multiple Praxis II area tests, or because they did their teacher training in North Carolina (where most institutions require the Praxis I). I use these multiple tests to construct a 'composite z-score' for teachers, which is simply the average of a teacher's z-scores on individual licensure tests.

In 1997, North Carolina's cut score on the Curriculum test was 153 while the cut score on the Content test was 127. In 2000, the state eliminated these individual test minimums and replaced them with a two-test combined cut score of 313. This change, coupled with the fact that applicants are allowed to take tests multiple times (each of the scores are reported in the teacher's record) and bank their scores, means that some teachers have relatively low reported scores for some tests.[13] I take advantage of this policy shift to address some concerns about sample selection, by identifying teachers in today's workforce who were hired under the 1997 cut scores but who would not have been granted entry under the 2000 cut scores and vice versa.

---

[11] I am grateful to the Educational Testing Service for supplying the test distribution information necessary to do this normalization.

[12] I could not obtain similar information for the NTE so I normalized the distribution to have a mean of zero and a standard deviation of one based on the performance of teachers in North Carolina (the sample in each year). An alternative is to norm the NTE based on the year in which teachers sat for the exam. I opted against this because I could only do this relative to the teachers who are in the North Carolina teacher workforce from 1994-95 to 2003-04 and there is clear evidence (discussed below) of nonrandom attrition from the teacher workforce.

[13] In addition, teachers may teach in a North Carolina school with a temporary license that is valid for one year without meeting the Praxis II requirement, but must then achieve an acceptable score on the Praxis II after this period in order to continue teaching. (North Carolina Department of Public Instruction, 2003). – where can you retrieve this now? If there is an url with the packet we should include that in the refs list - cw

In addition to this, I focus on the pass/fail signal associated with the established cut scores in Connecticut, a state in which many North Carolina teachers would be ineligible to teach (based on their existing reported Praxis scores).

The student achievement measures in the data come from state-mandated standardized reading and math tests that are tied to the North Carolina *Standard Course of Study*. These criterion-referenced tests are vertically aligned and used by the NCDPI's Accountability Department to determine performance and growth/gain goals and ratings for all schools as part of the state's "ABC" education reform program. All student test scores are normalized within grade and year to have a mean of zero and a standard deviation of one, so the coefficient estimates from the models described above measure the predicted impact of the independent variables on a student's standing within the performance distribution (in standard deviation terms).

The analysis is restricted to teachers who reported teaching a "self contained" class and those who have valid NTE and/or Praxis test scores.[14] It is also restricted to students in grades four through six who have a valid math and/or reading pre- and post-test score (for example, the end-of-year fourth-grade math score would be used as the post-test when a student's end-of-year third-grade math score was used as the pre-test).[15] The argument for these restrictions is that they allow an analysis for a group of students who are highly likely to be matched to their teachers of

[14] While I do not report these findings, I have experimented with various other samples, including using a larger sample by relaxing the requirement that teachers be in a self-contained classroom and instead matching teachers to students based on the subject code of class (i.e. that a teacher is reported as teaching a reading or math class. The findings discussed in the next section are not materially affected by these variations in the sample used for the analysis.

[15] The great majority of 6th grade students in North Carolina are enrolled in middle schools so the restriction of the sample to include only self-contained classes eliminates most 6th graders (they make up, on average, 1.9% of the student observations used in the analysis). For a discussion of the changes to a sample resulting from the restriction to self-contained classrooms, see Clotfelter et al. (2001).

math and reading.[16] This yields a sample of more than 23,740 unique teachers (69,025 teacher observations) and over 701,121 unique students (1,136,420 teacher-student observations).

Table 1 reports sample statistics for select variables by teacher licensure-test performance quintile.[17] The means reported in this table are derived by first averaging the class characteristics (class size, student achievement, etc.) for each teacher and then averaging across teacher observations, thus they should be interpreted as the characteristics faced by the average teacher within a given licensure performance classification.

There appear to be significant differences in both the characteristics of teachers who fall into different quintiles and in the students they teach. In particular, teachers who fall in the lower quintiles of licensure performance are much more likely to be minority teachers and they tend to be teaching more minority and disadvantaged students. There are no clear trends, however, for student performance; for instance, it is in the middle of the licensure performance distribution where we see students with the highest math and reading scores, and in quintile 4 where we see students who make the greatest gains in achievement.[18]

V. Results

In this section I describe the findings on the value of using teacher licensure tests as a screening mechanism and as a signal of teacher quality. Following that, I explore the potential sources of bias resulting from sample selection, the nonrandom match of teachers to schools and classrooms, or to nonrandom teacher attrition from the labor force. However a few peripheral findings warrant brief notice. In both reading and math and across the model specifications

---

[16] Teachers and students are matched based on the teacher of record listed on a student's state test. Students are not tested prior to the 3rd grade and they almost always switch teachers for grades above 6th.
[17] The sample is divided based on the teacher "composite z-score."
[18] Student test-score variables for post-test, pre-test, and growth are based on the average scores (post-test, pre-test, and growth) in that subject for students of a particular teacher in a particular year.

reported below, minority students (black and Hispanic), male students, participants in the free and reduced-price lunch program, those whose parents have less education, and/or those with reported learning disabilities score lower than their reference groups. Consistent with much of the educational productivity literature (for example, Hanushek, 1986; 1997), there is little evidence that a teacher having a masters degree (or higher) is a signal of teacher effectiveness. The findings also suggest that students of teachers who graduate from a North Carolina-approved training program outperform those whose teachers do not (that is, those who get a degree from an alternative state program or a program from outside the state) by about 1 percent of a standard deviation, and, consistent with recent evidence that NBPTS certification serves as a signal of quality (Cavalluzzo, 2004; Goldhaber and Anthony, forthcoming), I find that teachers certified by the National Board for Professional Teaching Standards (NBPTS) outperform non-certified teachers by 1 to 4 percent of a standard deviation, with larger effects in math. By contrast, there is only spotty evidence in the student math-achievement models, and none in the reading-achievement models, that graduating from a more selective college (based on the average institutional SAT of the college from which teachers graduated) leads to increased student achievement and the effects are consistently small.

Finally, I find that teachers see the largest gains in productivity during the early years of their career. Students with a teacher who has 1 to 2 years of experience outperform students with novice teachers by 3 to 7 percent of a standard deviation. Students with teachers who have 3 to 5 years of experience tend to outperform those with 1 to 2 years of experience (the difference is not statistically significant across all specifications) by about an additional 2 percent of a standard deviation; and students of teachers with 3 to 5 years of experience outperformed those

with novice teachers by 4 to 9 percent of a standard deviation.[19] I find little evidence, however, of statistically significant productivity gains associated with increases in experience beyond 5 years. These findings on teacher experience are broadly similar to those reported in Boyd et al. (2005), Clotfelter et al. (forthcoming), and Rivkin et al. (2005).

A. Licensure Tests as a Screening Device

Table 2 reports the estimated relationship between student achievement and whether a teacher passes or fails the licensure exam based on a set state standard.[20] These models correspond to equation 1 from Section III, and include an unusually rich set of student background controls (the specific independent variables used in each model specification are reported in notes below the table), but they exclude any teacher variables as I am interested only in assessing the pass/fail screening value of the test. The coefficients change very little, however, when the models include additional teacher controls.[21]

I begin (in column 1 for reading and 4 for math) by exploring whether North Carolina's current cut score (recall that this is a combined score of 313 on the Praxis II tests) serves as a signal of teacher quality. In these pass/fail models I focus on the sub-sample of teachers for whom the Praxis II (the required state test in North Carolina) information is available. Because teachers may teach in a North Carolina school with a temporary license that is valid for one school year without meeting the Praxis II requirement, there are some teachers who have not met the current testing requirement. There are also teachers who took the Praxis tests prior to 2000 whose scores would have made them ineligible to teach based on the new 2000 standard. Nearly

[19] The differences between 1 to 2 years of experience and 3 to 5 years are statistically significant.
[20] The standard errors reported in this and the following tables are corrected for clustering of students within classrooms.
[21] Recall the argument for this is that states do not require information on the aforementioned characteristics, so they should not be included in the models that identify the value of the pass/fail licensure test signal. The effect of these teacher characteristics on student achievement is captured by the licensure test variables (as well as the other variables included in the model) though their partial correlations.

a thousand teachers (for whom Praxis II information exists) are teaching in the state without having met the state's 2000 licensure test eligibility requirements.[22] The positive coefficient estimates on the pass/fail indicator variable, which provides a comparison of the teachers who did not meet the 2000 standard to those who did, suggest there is a value to using licensure tests as a screen: teachers who meet the current North Carolina standard are more effective in math than those who do not, by about 6 percent of a standard deviation. The findings for reading are smaller (about 2 percent of a standard deviation) and only marginally significant (at the 12 percent level).

As I described in Section III, there is reason to believe that sample selection could bias the estimated impact of passing the licensure test relative to failing it. In particular, I would hypothesize a downward bias in the estimate of the North Carolina pass/fail coefficients as one might imagine that teachers who are in the profession, despite not meeting the standard, have unobserved attributes that make them effective in the classroom. While it does not directly address this potential source of bias, an interesting question to consider is the extent to which the pass/fail coefficient changes if a different cut score is utilized. For example, if I arbitrarily declare a much higher cutoff, suddenly more "failures" will appear in the sample. The test's developers (ETS), however, caution against such arbitrariness, saying that scores other than the cut score may not have the same predictive value (for instance, because of increased measurement error away from the cut point).

But what about another state's cut score, or a previous cut score in North Carolina? Surely state education policymakers do not consider their cut scores to be arbitrary? North

[22] 389 teachers in the sample who have Praxis II test information on their records did not meet the 2000 standard but are teaching under a temporary license. An additional 552 teachers met the 1997 state standard but, based on their reported scores, would not have met the 2000 test standard.

Carolina's earlier (1997-2000) standard offers one interesting comparison, and Connecticut's pass/fail standard offers another. While Connecticut's standard isn't directly comparable to the current North Carolina combined cut score, it is directly comparable to North Carolina's earlier two-test standard. Both states had (and Connecticut still has) cutoff requirements on the same two sub-sections of the Praxis II test. The Connecticut requirements, however, were about .6 standard deviations higher than North Carolina's 1997 requirement on the Praxis II Curriculum, Assessment, and Instruction and 1.75 standard deviations higher than their Praxis II Content Area Exercises requirement.[23] Consequently, a significant portion of the North Carolina teacher sample would not be eligible to teach based on the Connecticut standard.[24] Given that Connecticut is using the assessment to determine teaching eligibility, one can reasonably assume that its standard is valid around its cut score.

Columns 2 (for reading) and 5 (for math) of **Table 2** report the results for the value of the pass/fail indicator for the 1997-2000 North Carolina pass/fail signal. A comparison of columns 1 and 2 (for reading) and 4 and 5 (for math) shows that the estimated value of the pass/fail signal does not change much when moving from the current North Carolina standards back to those utilized between 1997-2000.[25] Given that the standard did not significantly change with North Carolina's adoption of the combination score requirement (getting a combined 313 on the two

---

[23] These differences are based on the national distributions on the Praxis II tests obtained from ETS. Connecticut requires a 163 on the Curriculum, Instruction, and Assessment test and a 148 on the Content Area test (for information on state requirements, see http://www.ets.org/portal/site/ets/menuitem.22f30af61d34e9c39a77b13bc3921509/?vgnextoid=d378197a484f4010VgnVCM10000022f95190RCRD).

[24] There are 3,341 teachers in North Carolina whose reported scores would have made them ineligible to teach in Connecticut based on the Connecticut Praxis cutoff requirements. Though the difference in combined scores is small (and Connecticut actually has a lower combined requirement), the fact that Connecticut requires minimum scores on both exams means that many teachers who just meet or slightly exceed the combined requirement in North Carolina will fail by Connecticut standards unless they happen to get the exact distribution of scores dictated by Connecticut's minimum scores for each test.

[25] There are teachers (389) in the workforce who did not meet the standard because of the ability to teach with a temporary license and, in addition, there are 61 teachers who met the state standard in 2000 (the combined score) but, based on their reported test scores, would not have met the state's 1997 standard.

Praxis tests) in 2000, it is not surprising that the coefficient estimates are quite similar to those for the current licensure standards, as there isn't a significant change in the standard or the sample of test failers.

In column 3 (for reading) and 6 (for math) of the table, I report the results of using a significantly higher cutoff standard. This analysis provides an unbiased indication of what the impact of the Connecticut test would have were the potential pool of teachers to be the North Carolina teacher workforce.[26] Contrary to what one would expect if it sample selection led to an downward bias in the estimated effects of the licensure test pass/fail coefficients, the coefficient on the higher Connecticut standard is actually quite a bit *smaller* than the estimates based on the lower North Carolina standard. The Connecticut standard appears to be a weaker signal of quality, at least as judged by the magnitudes of the pass/fail coefficients. The finding that raising the bar to the Connecticut standard reduces the magnitude of the pass/fail coefficient results from the fact that a significant proportion of the sample gets reclassified, about 7.5 percent; this reclassification increases the estimated average teacher contribution to student achievement of failing teachers while having almost no impact on the estimate of the average contribution to student achievement of those teachers who are in the passing category.

B. Licensure Tests as a Signal of Teacher Quality

The signal value of licensure test performance for teachers in the workforce is an important policy issue, as local school districts might wish to use this signal in helping to make hiring decisions. To investigate this, I estimate student achievement models (corresponding to equation 2) that focus on the entire teacher test-performance distribution, using model

[26] It is worth noting that the Connecticut and North Carolina teacher workforces differ in many ways. For example, Connecticut tends to have higher pay scales and has significantly lower attrition from the teacher labor force (4.8% annually) than does North Carolina (8.4% annually) (http://www.all4ed.org/press/pr_081505.html).

specifications that include dummy variables that indicate the quintile of teacher performance on the test (the lowest quintile being the reference category).[27] Unlike the pass/fail models discussed above, these models include a full set of teacher variables to represent the fact that districts have more information about teachers during the hiring process and likely care about the information value of tests beyond that provided by the other teacher variables.

Table 3 presents the estimated coefficients for several different tests that are available on teacher records. Recall that there are two distinct Praxis II tests currently required by the state: the Praxis II Curriculum test and the Praxis II Content test. Columns 1 and 3 report the estimated coefficients of student reading and math achievement models that include teacher licensure performance on the Content test, and Columns 2 and 4 reported these estimated coefficients for the Curriculum test. Casual observation suggests there is relatively little evidence that the Praxis II Content test predicts student achievement in either math or reading. The teacher Content test is only statistically significant for the top quintile of performance (relative to the bottom quintile) in math. By contrast, there is far more evidence that the Praxis II Curriculum test provides a signal of teacher effectiveness. While not all of the quintile coefficients are statistically significant, the point estimates in reading suggest a consistent positive relationship between teacher performance on the test and student achievement, and F-tests confirm that the null hypothesis of no significance can be rejected for both students' math and reading achievement. The coefficient estimates of the two teacher tests are surprisingly similar when both tests are entered into the same model (they only change in the thousandths place), and an F-test shows the coefficients on

[27] While not reported here, models that allow for a linear relationship between the teachers' composite z-scores and student achievement suggest a small positive relationship between the two. Specifically, the point estimates suggest that a 1 standard deviation change in teacher test-score performance is predicted to increase student test scores by slightly less than 1 percent of a standard deviation in reading and slightly less than 2 percent of a standard deviation in math. These results are along the same order of magnitude as those of Clotfelter et al. (forthcoming) and Goldhaber (2005).

the curriculum test quintiles remain marginally statistically significant (above the 90th percent confidence level in both math and reading).

Teachers who score in the top quintiles on the Curriculum test appear to be significantly more effective: relative to the bottom quintile, teachers in quintile 4 have students who score about 1.7 percent of a standard deviation higher and those who are in the top quintile score about 2.4 percent of a standard deviation higher (the difference between the two top quintiles is not statistically significant). In the math models, the only significant difference between quintiles is between the top quintile and the bottom. The point estimate suggests that students of teachers scoring in the top quintile would have math achievement scores that are over 3 percent of a standard deviation higher than those whose teachers score in the bottom quintile. And, while the other coefficients are not statistically significant at the 5 percent confidence level, the pattern of results (the upward trend between teacher performance and teacher effectiveness) is consistent with the findings in the reading models.[28]

When I expand the sample to the 'composite z-score sample', which includes teachers who have *any* licensure test on their record (and therefore many more experienced teachers) the findings are even stronger.[29] The findings from these model specifications are reported in columns 2 (for reading) and 4 (for math). In reading, there is a clear upward trend between teacher licensure performance and student achievement. Teachers in quintile 2 are estimated to produce student achievement gains that are 1.5 percent of a standard deviation higher than those of teachers who score in the lowest quintile. Teachers in quintiles 3 and 4 are estimated to

[28] Many teachers in the state also have a Praxis I test score on their records, because this test is required by most in-state teacher training institutions. While I do not report the results in this table, there is evidence that teachers who do well on the math component of the test are more effective in teaching math.

[29] The use of the composite z-score allows me to include in the models information from the full sample of teachers (for whom any licensure exam score—NTE, Praxis I or Praxis II—exists on a teacher's record), and not just teachers hired after 1996 at the time the Praxis test came into use.

produce student achievement gains that are just over 2 percent of a standard deviation higher (the difference between these quintiles is not statistically significant), and those in quintile 5 are estimated to produce student achievement gains that are over 3 percent of a standard deviation higher than teachers at the bottom (the difference between quintiles 5 and either quintiles 3 or 4 is statistically significant).

While I do not report these findings, I also estimated model specifications exploring whether the signal value of licensure tests may be different for different types of teachers or students. There is, for instance, some evidence of test "prediction bias" (when a test differentially predicts performance, for example, college or job performance) for individuals from different race/ethnicities (Vars and Bowen, 1998).[30] In general the pattern of results reported in **Table 3** is found for both black and white teachers and for the various sub-groups of students: students who have higher-performing teachers tend to have higher achievement levels in both reading and math. In particular, it does not appear that racial prediction bias exists in the context of licensure testing, as models that allow the licensure test-student achievement relationship to differ by teacher race (for example, by including race licensure quintile interaction terms) show little evidence that it does so. There are also some interesting differences in findings across student types: for example, the impact of having a higher-scoring teacher tends to be larger for non-black students and those who are not eligible for free and reduced-price lunch.[31]

---

[30] The SAT test, for example, tends to over-predict the academic performance for black college students (Jencks, 1998), empirical evidence dating from the *Coleman Report* (Coleman, 1966) tends to find that teacher quality has a larger impact on lower-achieving students than on those who are higher-achieving, and there is some evidence (Dee, 2004) of role-model effects in education, suggesting that teacher effects may vary depending on whether teachers and students are matched based on race/ethnicity.

[31] There is also some evidence, particularly in math, of the type of role-model effects found by Dee (2004). For example, while not consistently statistically significant, the magnitude of the estimates suggest that black teachers who are teaching black students tend to outperform black teachers who are teaching white students.

Thus, on the whole, the findings suggest that teacher test performance, particularly on the Curriculum test, does provide a signal of teacher effectiveness across the performance distribution for different types of teachers teaching different types of students; however, as I noted in Section III, there are at least three sources of bias that threaten the assessment of the licensure test signal. I examine these potential sources of bias below.

C. Threats to the Validity of the Measure of the Teacher Test Signal

Sample Selection

The potential for bias arising from sample selection is a difficult problem to deal with in this context, because teachers who are not in the workforce are not observed. That said, there are some compelling reasons to believe that sample selection is not a significant problem. First, empirical evidence suggests that relatively few teachers (no more than 10-15 percent) are screened out of the labor market by licensure tests (Angrist and Guyan, 2004). Second, a significant share of the sample of North Carolina teachers used here actually failed the North Carolina standard because of the shifts in state policy. Third, I utilize the fact that a considerable number of teachers in the sample (17 percent) have multiple scores for the same test because they initially failed to achieve the required cutoff. With this information, I estimate the pass/fail models based on a teacher's lowest reported score to see whether teachers who initially fail a test but pass on a re-take appear to be systematically different from those who pass on their first attempt. The estimated coefficients from models that utilize a teacher's lowest reported score are not appreciably different from those reported in the models specified in **Table 3**. Furthermore, there is no evidence from models that include a dummy variable indicating an initial (or multiple) test failure that teachers who re-take a test are systematically different from those who do not.

Finally, rather than increasing the strength of the pass/fail signal, the shift from the North Carolina standard to the higher Connecticut standard actually lowered it. Were it the case that individuals in the teacher labor market with failing scores tended to have unobservable attributes positively correlated with student achievement, one would expect the increase in the standard to increase the magnitude and significance of the pass/fail indicator variable, and a smaller percentage of the teacher observations falling into the failing category would have positive errors. This last point is buttressed by a closer examination of the signal value of the tests along the distribution.

Bias in the signal value of the teacher licensure test is most likely due to those who are in the labor market despite having failed to achieve the required cutoff score. Thus, one might expect the signal value associated with linear changes in the score to be different for these teachers. To test this, I employ a quasi regression discontinuity approach in estimating a model where the licensure score is entered as a linear explanatory variable, but which also allows the slope and intercept of the licensure score coefficient to vary above and below the cutoff score. Further, in this model I restrict the sample to teachers who have entered the labor force since 2000, since this is a sample of teachers who clearly entered the teacher labor market despite having failed to achieve the post-2000 North Carolina test cutoff requirement. An F-test fails to reject the hypothesis that the slope and intercept of the licensure test for failing teachers is different than it is for passing teachers, thus suggesting no bias exits. While all the above specification tests are imperfect ways of assessing the implications of having mainly test-eligible teachers in the sample, none of the tests suggests that this type of sample selection leads to systematic bias in the estimates of the pass/fail signal.

Nonrandom Teacher Sorting

A second threat to the validity of the measure of the teacher test signal is that the estimates of the teacher test coefficients are biased due to nonrandom matching of teachers and students: there is evidence that teachers tend to sort across students such that the more-senior, credentialed teachers are teaching the higher-achieving students (Lankford et al. 2002), and that this sorting pattern affects the estimated returns to teacher characteristics (Clotfelter et al., forthcoming; Goldhaber and Anthony, forthcoming). I address this issue by estimating models that include school and student fixed effects, which are reported in **Table 4**. In these specifications, the teacher effects are identified based on variation in teacher qualifications within schools across classrooms (in the case of the school fixed-effects model) and across students over time (in the case of the student fixed-effects model).[32] I do not estimate model specifications that include *both* school and student fixed effects, as models would be identified solely by students who switch schools.

In these model specifications I utilize the 'composite z-score sample'. For comparison purposes, the coefficient estimates for models that do not include fixed effects are reported in columns 1 (for reading) and 4 (for math). Columns 2 and 5 report the specifications that include school fixed effects, and columns 3 and 6 report the specifications that include student fixed effects. Focusing on the full sample (the novice teacher sample is discussed below) reported in Panel A of Table 4 illustrates that school and student fixed effects are significant predictors of student achievement in both reading and math models.[33] And the results of these models suggest that teacher sorting accounts for some of the positive correlation between teacher and student test performance. The comparison between similar models that do not include fixed effects (columns

[32] A lagged measure of student achievement *is not* included in these models, though the inclusion of this lag does not appreciably change the coefficients of the teacher licensure test variables.
[33] The p values for all F-tests fall below 0.01.

1 and 4 for reading and math, respectively) with the school fixed-effects models (columns 2 and 5, respectively) or with the student fixed-effects models (columns 3 and 6, respectively) shows a marked decrease in the point estimates for the teacher test performance quintile variables.

In reading, the coefficient estimates drop by around half (for each quintile) when moving from no fixed effects to school fixed effects. The decrease in the magnitudes of the quintiles is smaller when moving from school to student fixed-effects models, and there are no longer statistically significant distinctions between the top four quintiles; however, the point estimates continue to suggest that teachers in any of the top four quintiles are more effective than those who score in the bottom quintile. In math, the magnitude of the teacher performance quintiles drops when moving from no fixed effects to school fixed effects to student fixed effects, but the drop in magnitude is far less, percentage wise, than in the reading models. And, in the math models there continues to be a statistically significant difference between those teachers at the top of the distribution and those below.

In sum, these results show that the nonrandom sorting of teachers across schools and students does have an impact on the estimated relationship between teacher test performance and student achievement; however, the findings continue to suggest that performance on these tests does provide a signal of teacher effectiveness. Finally, results from fixed-effects specifications of the teacher pass/fail models are not materially different from the estimates reported in **Table 2.** The coefficient estimate from the reading pass/fail model with school or student fixed effects is not statistically significant, as was the case before. The coefficient estimate from the math pass/fail model with fixed effects is smaller (suggesting effects in the range of 3 to 4 percent of a standard deviation), but still statistically significant. Thus, it does appear that, at least for math achievement, the state cutoff represents an important teacher-quality screen.

Nonrandom Teacher Attrition

To address the final identified threat to the assessment of the validity of the licensure test signal—nonrandom attrition from the sample—I focus on a sample of novice (first-year) teachers, replicating the model specifications reported above for Panel A of **Table 4**.[34] The estimated coefficients from these novice teacher models, without fixed effects (columns 1 and 4), with school fixed effects (columns 2 and 5), and with student fixed effects (columns 3 and 6) are reported in Panel B **Table 4**.

The findings from models with no fixed effects as well as those with school fixed effects tends to confirm those reported earlier: teacher licensure test performance *is* a signal of teacher effectiveness. If anything, these results suggest a stronger signal of licensure test performance: outside of the student fixed-effects models, the magnitudes of the coefficients are consistently larger in the novice sub-sample than the full teacher sample (a comparison of estimates across Panels A and B) and this is true for both reading and math achievement models. This is, in fact, what one would expect if the teachers who do well on the licensure exam and leave the North Carolina teacher workforce are the same ones who tend to produce a high level of value-added for students.

The story is quite different for the student fixed-effects models, as few of the coefficients are statistically significant for either math or reading achievement. I interpret these findings cautiously, however—these models are identified by students who have at least two consecutive novice teachers, and this sample of students is relatively small (3,900). Since student assignment is clearly not random, with low-achieving students tending to be assigned to less-experienced and less-credentialed teachers (Lankford et al., 2002), there is some concern that students who

[34] This sample includes 5,607 unique teachers and 88,703 unique students.

are repeatedly assigned to novice teachers are systematically different from other students in ways that are not accounted for by the time-invariant student effects.

V. Public Policy Implications and Conclusions

Licensure testing as a requirement for employment in public schools is widespread, despite the fact that there is little quantitative research showing its efficacy. The results presented here generally support the hypothesis that licensure tests are predictive of teacher effectiveness, particularly in teaching mathematics, and this finding is robust to alternative specifications of the model, including those that account for nonrandom sorting of teachers across students. If states are seeking criteria to ensure a basic level of quality, then licensure tests appear to have some student achievement validity.

What do the results mean in terms of student achievement? That depends on how teacher tests are used in shaping the teacher workforce. Their most straightforward use is in determining employment eligibility, and the point estimates from North Carolina's pass/fail cutoffs suggest that teachers who pass the test produce, on average, student achievement gains that are in the range of 3 to 6 percent of a standard deviation higher (in math) than those who fail.[35] The approach usually employed to interpret the magnitude of an educational intervention is to measure the effect size relative to the standard deviation of the test (either the level, as is done here, or the gain). Recent work, however, suggests that this standard may understate the true impact of the intervention because of measurement error inherent in any type of testing (Boyd, Grossman, Lankford, Loeb, and Wyckoff, 2006).[36]

---

[35] Based on the findings from column 4 of Table 3 and similar models (not reported) that include either school or student fixed effects.

[36] This study shows that measured effects relative to estimates of the standard deviation of the *universal* score (an estimate of the underlying academic achievement, purged of test measurement error) substantially understate the effects of educational resources when the effects are measured relative to the standard deviation of student *gain*.

An alternative way to gage the effect is to judge it against the impact of other educational resources. Recall that teachers are estimated to become more effective with experience: students of teachers with 1 to 2 years of experience outperform students of novice teachers by 5 to 7 percent of a standard deviation. This suggests, very roughly, that the average teacher who fails to achieve the licensure standard on the test, were they allowed to teach, would be anticipated to produce the same level of student math achievement in her second or third year as a novice teacher who did achieve the state standard on the licensure test.

While it is interesting to look at the *average* effects of teacher testing as a signal, it is also informative to determine the distributional consequences of using these tests to determine employment eligibility. It is here that it becomes clear that teacher testing is not without its costs. Specifically, because the point estimates do not provide evidence of a terribly strong relationship between teacher test performance and student achievement, there are likely to be a significant number of false negatives (individuals who fail to achieve a minimum requirement on the licensure test but who would have been high-quality teachers) and false positives (individuals who do well on the licensure test but who are not very effective teachers).

Figure 1, which shows the scatterplot of estimated value-added teacher effects in mathematics plotted against the Praxis II Curriculum test (more heavily shaded circles represent a greater population concentration), offers a final illustration of these tradeoffs.[37] For the sake of illustration, I have defined a minimum level of teacher quality (TQ_{min}) to be 2 standard

---

scores. In this case, the findings suggest that the true impacts of educational resources may be as much as four times larger. However, the degree to which effects are understated when using the standard deviation in test score *levels* is far lower, on the order of 25 percent.

[37] The value-added teacher effects are based on a sample of teachers who have a valid Curriculum test score on their records and a model similar to that presented in column 5 of Table 2. The only difference between this specification and that used to generate the teacher effects is that all individual teacher variables are excluded from the model and replaced with a teacher fixed effect. The fixed effect is estimated based on teacher observations across years, then standardized to have a mean of zero and a standard deviation of one. The results from a similar analysis of student reading achievement yield qualitatively similar findings.

deviations below the median teacher effectiveness (based on the standardization of teacher effects, 97.5 percent of teachers will fall above this quality threshold). I use the 1997-2000 North Carolina standard on the Content Area portion of the Praxis II because it can easily be compared to the Connecticut standard on the same test. There are a number of false positives, represented by the sum of areas VIII and IX (about 2 percent of the teacher workforce), and of false negatives, represented by the sum of areas I and IV (just over 3 percent of the teacher workforce) under the 1997 North Carolina standard. Interestingly, a high proportion of the false negatives, about 36 percent, actually had estimated teacher effects that exceed the estimate of the mean teacher effect (those in area I). Were the testing standard to be raised to the higher one utilized in Connecticut, the number of false positives would fall in the area represented by VIII, but the trade-off is that we see a very large increase in the number of false negatives, represented by the sum of areas II and V (about 7 percent of the workforce).[38] Recall that the estimates suggest a stronger relationship between teacher performance on the Curriculum test and student achievement than the relationship between teacher performance on the Content test and student achievement; since Figure 1 is based on the test that is more strongly correlated with student achievement, it provides a conservative measure of the number of false positives and negatives (comparable analyses of the Content test show a much higher proportion of false positives and negatives).

Whether the above trade-offs are worthwhile is a value judgment, and likely depends both on labor market conditions (for example, how difficult it is to hire teachers) and the harm that low-quality teachers might do to students. But, it is also important to consider the role local

[38] Variation in the teacher test cutoff standard shows that the probability that a teacher in the labor force has an impact larger than the arbitrary standard of two standard deviations below the mean is maximized with a cut score of 150 on the Curriculum test (slightly lower than the North Carolina standard).

school systems play in determining the quality of the teacher workforce. For example, it is not clear that potential teachers screened out by a licensure test would be in the teacher workforce simply by virtue of being eligible to teach, as the link between the eligible pool of teachers and the teacher workforce depends on school system hiring; thus it is difficult to say what the impact of any change in the testing standard might be. There is relatively little quantitative evidence on the selection processes of school systems, and none (that that I am aware of) that focuses on whether or not the information from licensure tests is used by local school systems in hiring decisions.[39] Nor is there much evidence on whether hurdles associated with licensure tests (or other licensure requirements for that matter) affect the number of people who opt to pursue a career in teaching. These missing links certainly mean that one should not expect any change in a state's licensure test cutoff score to result in a comparable change in the licensure test performance of newly hired teachers.

Teaching is not alone among professions in requiring candidates to meet a minimum test competency. The bar exam and medical boards are correlates to licensure tests, but there are also many examples of professions that test their prospective employees without using a cut score as an absolute determining factor for employment eligibility; college professors, for instance, can practice without satisfying any absolute test standard. The research presented here suggests that licensure test performance is clearly not a 'silver bullet' credential that can be used to predict teacher effectiveness. If anything, this speaks to the need for districts to be selective when hiring teachers. A large body of empirical evidence suggests that credentials like teacher licensure provide only a weak signal of teacher quality. To the extent that hiring officials cannot a priori discern more subtle teacher attributes that predict effectiveness, they must consider policies

[39] For information on school system hiring, see Ballou (1996); Wise et al. (1987), Guarino et al. (2004); Liu (2005); or Liu and Quinn (2005-2003 in refs list – pls check).

designed to shape their labor forces once they have had the chance to observe the effects of teachers in the classroom.

REFERENCES

Angoff, W. H. (1971). Scales, Norms, and Equivalent Scores. Educational measurement (by) William H. Angoff (and others). R. L. Thorndike, 1910-. Washington, American Council on Education: xx, 768 p.

Angrist, J. D. and J. Guryan (2004). "Teacher Testing, Teacher Education, and Teacher Characteristics." American Economic Review 94(2): 241-46.

Ballou, D. (1996). "Do Public Schools Hire the Best Applicants?" Quarterly Journal of Economics 111(1): 97-133.

Boyd, D., P. Grossman, H. Lankford, S. Loeb, and J. Wyckoff. (2005). How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement: 41.

Boyd, D., P. Grossman, et al. (2006). The Assessment of Effect Sizes in Value-Added Analysis: Accounting for Test Measurement Error, Teacher Policy Research.- I can't find this paper. Who are the et al authors?-cw

Cavalluzzo, L. C. (2004). Is National Board Certification an Effective Signal of Teacher Quality? The CNA Corporation. 2006.

Clotfelter et al. (2001).

Clotfelter, C. T., H. F. Ladd, and J. Vigdor. (Forthcoming). Teacher Sorting, Teacher Shopping, and the Assessment of Teacher Effectiveness. Journal of Human Resources.

Coleman, J. S. (1966). Equality of educational opportunity. Washington, DC, U.S. Dept. of Health, Education, and Welfare, Office of Education.

Dee, T. S. (2004). "Teachers, Race, and Student Achievement in a Randomized Experiment." Review of Economics and Statistics 86(1): 195-210.

Educational Testing Service (2006). Proper Use of the Praxis Series and Related Assessments. 2006.

Ehrenberg, R.G. and D.J. Brewer (1995). "Did Teacher Verbal Ability and Race Matter in the 1960s? Coleman Revisited" Economics of Education Review, 14(1): 1-21.

Ehrenberg, R. G., D. D. Goldhaber, et al. (1995). "Do Teachers Race, Gender, and Ethnicity Matter - Evidence from the National Educational Longitudinal-Study of 1988." Industrial and Labor Relations Review 48(3): 547-561.

Ferguson, R. F. (1991). "Paying for Public-Education - New Evidence on How and Why Money Matters." Harvard Journal on Legislation 28(2): 465-498.

Ferguson, R. F. (1998). Teachers' perceptions and expectations and the Black-White test score gap. The Black-White test score gap. C. Jencks and M. Phillips. Washington, DC, Brookings Institution Press: 273-317.

Ferguson, R. F. and H. F. Ladd (1996). How and why money matters: An analysis of Alabama schools. Holding schools accountable: performance-based reform in education. H. F. Ladd. Washington, DC, The Brookings Institution.

Flippo, R. F. (2002). "Repeating History: Teacher Licensure Testing in Massachusetts." Journal of Personnel Evaluation in Education 16(3): 211-29.

Friedman and Kuznets (1945)

Goldhaber, D. (2004). Why Do We License Teachers? A Qualified Teacher in Every Classroom: Appraising Old Answers and New Ideas. F. Hess, A. J. Rotherham and K. Walsh. Cambridge, MA, Harvard Education Press: 81-100.
Goldhaber, D. (Forthcoming). Teacher Licensure Tests and Student Achievement: Is Teacher Testing an Effective Policy? Learning from Longitudinal Data in Education. D. Chaplin and J. Hannaway. Washington, DC, UI Press.
Goldhaber, D. and E. Anthony (Forthcoming). "Can Teacher Quality be Effectively Assessed? National Board Certification as a signal of Effective Teaching." Review of Economics and Statistics.
Guarino, C., L. Santibanez, G. Daley, and D. Brewer. (2004). A Review of the Research Literature on Teacher Recruitment and Retention. Santa Monica, CA, RAND Corporation.
Hanushek, E. A. (1986). "The Economics of Schooling - Production and Efficiency in Public-Schools." Journal of Economic Literature 24(3): 1141-1178.
Hanushek, E. A. (1992). "The Trade-off Between Child Quantity and Quality." Journal of Political Economy 100(1): 84-117.
Hanushek, E. A. (1997). "Assessing the Effects of School Resources on Student Performance: An Update." Educational Evaluation and Policy Analysis 19(2): 141-64.
Hanushek, E. A. and R. R. Pace (1995). "Who Chooses to Teach (and Why)?" Economics of Education Review 14(2): 101-17.
Hanushek, E. A., S. G. Rivkin, and L.L. Taylor. (1996). "Aggregation and the estimated effects of school resources." Review of Economics and Statistics 78(4): 611-627.
Jaeger, R. M. (1999). Some Psychometric Criteria for Judging the Quality of Teacher Certification Tests, Center for Education, National Research Council.
Jencks, C. (1998). Racial Bias in Testing. The Black-White Test Score Gap. C. Jencks and M. Phillips. Washington DC, Brookings Institution Press: 55-85.
Kane, T., J. Rockoff, et al. (2006). What Does Teacher Certification Tell Us About Teacher Effectiveness? Evidence from New York City.
Killingsworth, M. R. (1983). Labor Supply. New York, Cambridge University Press.
Kleiner, M. M. (2000). "Occupational licensing." Journal of Economic Perspectives 14(4): 189-202.
Lankford, H., S. Loeb, and J. Wyckoff. (2002). “Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis.” Educational Evaluation and Policy Analysis 24 (1):38-62.
Liu, E. (2005). Hiring, Job Satisfaction, and the Fit Between New Teachers and Their Schools. American Educational Research Association. Montreal, Quebec.
Liu, E. and M. Quinn (2003). Missed Opportunities: How We Keep High-Quality Teachers Out of Urban Classrooms. New York, The New Teacher Project.
Loeb, S. (2001). Teacher quality: Its enhancement and potential for improving pupil achievement. Improving Educational Productivity. D. Monk, H. Walberg and M. Wang. Greenwich, CT, Information Age Publishing: 99-114.
Mitchell, K., D. Robinson, et al., Eds. (2001). Testing Teacher Candidates: The Role of Licensure Tests in Improving Teacher Quality. Washington DC, National Academy Press.
Murnane, R. J. and R. J. Olsen (1990). "The Effects of Salaries and Opportunity Costs on Length of Stay in Teaching: Evidence from North Carolina." Journal of Human Resources 25(1): 106-24.
National Research Council. (2001).

North Carolina Department of Public Instruction. (2003). North Carolina Teacher Licensure Application Packet. 2006.
Nye, B., S. Konstantopoulos, et al. (2004). "How large are teacher effects?" Educational Evaluation and Policy Analysis 26(3): 237-257.
Podgursky, M., R. Monroe, and D. Watson. (2004). "The Academic Quality of Public School Teachers: An Analysis of Entry and Exit Behavior." Economics of Education Review 23(5): 507-518.
Proper Use of the Praxis Series and Related Assessments, 2006
Rivkin, S., E. A. Hanushek, and J.F. Kain. (2005). "Teachers, Schools and Academic Achievement." Econometrica 73(2): 417-458.
Rockoff, J. E. (2004). "The Impact of Individual Teachers on Students Achievement: Evidence from Panel Data." American Economic Review 94(2): 247-252.
Rotherham, A. J. and S. Mead. (2004). Back to the Future: The History and Politics of State Teacher Licensure and Certification. In A Qualified Teacher in Every Classroom? Appraising Old Answers and New Ideas. F. Hess, A. J. Rotherham and K. Walsh. Cambridge, MA, Harvard Education Press: 11-47.
Sack, J. L. (2005). NCATE Approves Single Cutoff Score on Teacher Tests. Education Week. Washington, DC. 25: 3-4.
Stigler (1971)
Stinebrickner, T. R. (2001). "A Dynamic Model of Teacher Labor Supply." Journal of Labor Economics 19(1): 196-230.
Stinebrickner, T. R. (2002).
Strauss, R. P. and E. A. Sawyer (1986). "Some new evidence on teacher and student competencies." Economics of Education Review 5(1): 41 - 48.
Summers and Wolfe (1975)
Title II Technical Assistance. (2003). October 2003 State Report. 2004.- check this info.
Vars, F. E. and W. G. Bowen (1998). Scholastic Aptitude Test Scores, Race, and Academic Performance in Selective Colleges and Universities. The Black-White Test Score Gap. C. Jencks and B. R. Phillips. Washington, D.C., Brookings Institution Press: 457-479.
Wise, A. E., L. Darling-Hammond, B. Berry, D.C. Berliner, E. Haller, A. Praskac, and P.C. Schlechty. (1987). Effective Teacher Selection: From Recruitment to Retention. Santa Monica, CA, RAND Corporation.

Table 1. Selected Sample Statistics by Teacher Licensure Test Score (Composite Z-score) Quintiles
(standard deviations in parentheses)

	Teacher Licensure Test Quintiles									
	Quintile 1		**Quintile 2**		**Quintile 3**		**Quintile 4**		**Quintile 5**	
	Student Test Scores									
	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math
Post-test	169.665	196.807	175.321	205.493	180.237	210.196	176.866	208.378	173.419	202.750
	(41.977)	(53.419)	(44.531)	(53.253)	(47.242)	(53.374)	(45.246)	(53.420)	(43.181)	(53.657)
Pre-test	153.282	176.464	156.868	184.139	161.238	189.683	157.585	186.867	156.339	181.430
	(31.018)	(51.660)	(33.779)	(53.139)	(38.109)	(54.664)	(33.896)	(53.728)	(32.262)	(52.468)
Growth	16.382	20.343	18.453	21.354	18.999	20.513	19.281	21.511	17.080	21.320
	(31.800)	(34.486)	(34.152)	(35.486)	(34.918)	(34.909)	(35.116)	(35.665)	(32.678)	(35.075)
	Teacher Characteristics									
Black	0.427		0.142		0.120		0.043		0.018	
	(0.495)		(0.349)		(0.325)		(0.202)		(0.134)	
White	0.548		0.845		0.869		0.949		0.976	
	(0.498)		(0.362)		(0.337)		(0.220)		(0.152)	
Female	0.938		0.922		0.902		0.914		0.933	
	(0.242)		(0.269)		(0.297)		(0.281)		(0.251)	
Years of teaching experience	15.978		11.871		12.242		10.540		12.405	
	(9.646)		(9.588)		(10.434)		(9.483)		(9.560)	
Masters Degree	0.262		0.237		0.241		0.264		0.337	
	(0.440)		(0.425)		(0.428)		(0.441)		(0.473)	
Fully Licensed	0.954		0.926		0.858		0.882		0.920	
	(0.209)		(0.261)		(0.349)		(0.323)		(0.272)	
	Class Characteristics									
Fraction of free/ reduced-price lunch students	0.608		0.535		0.529		0.496		0.475	
	(0.319)		(0.324)		(0.321)		(0.319)		(0.319)	
Fraction of minority students	0.512		0.415		0.423		0.376		0.361	
	(0.353)		(0.336)		(0.334)		(0.321)		(0.313)	
Class size	16.081		16.527		16.593		16.907		17.099	
	(8.572)		(8.553)		(8.398)		(8.461)		(8.339)	
Sample Size	14,022	14,071	14,022	14,072	14,021	14,069	14,022	14,073	14,021	14,068

Note: Sample sizes reflect multiple observations per teacher. Descriptive statistics for teacher and class characteristics use the math sample in reporting, as the math sample is slightly larger.

Table 2. Relationship Between State-Imposed Teacher Test Score Requirements and Student Achievement

(robust standard errors in parentheses)

	Reading			Math		
	1	**2**	**3**	**4**	**5**	**6**
Pass/Fail the Exam	0.02 (0.011)	0.026 (0.018)	0.001 (0.006)	0.055** (0.015)	0.069** (0.026)	0.012 (0.008)
Licensure Exam Standard	Current NC Standard	1996 NC Standard	Current CT Standard	Current NC Standard	1996 NC Standard	Current CT Standard
R2	0.66	0.66	0.66	0.7	0.7	0.7
Sample Size	173,522	173,522	173,522	174,589	174,589	174,589

**, *: Significant at 1% and 5% confidence level, respectively.

Note: Sample sizes reflect student-teacher observations. In addition to the variables specified above, the models also include the following controls: grade-level, year, a student's pre-test score, grade, race/ethnicity, gender, free/reduced-price lunch status (free/reduced-price lunch information is available for 1999 and beyond), parental education (whether a parent has a BA), Limited English Proficiency status, learning disability status, class size, and percent minority in the class. Mean value replacement was used for cases where values for the explanatory variables were missing.

Table 3. Licensure Test Performance as a Signal

Robust standard errors in parentheses

	Reading			Math		
	1	2	3	4	5	6
	Content Test	Curriculum Test		Content Test	Curriculum Test	
Teacher Licensure Performance Along the Test Distribution (reference group is bottom quintile)						
Quintile 2	0.003 (0.007)	0.004 (0.008)	0.015** (0.004)	0.001 (0.010)	0.004 (0.012)	0.024** (0.006)
Quintile 3	-0.002 (0.007)	0.007 (0.008)	0.021** (0.004)	0.005 (0.011)	0.018 (0.012)	0.028** (0.006)
Quintile 4	0.004 (0.007)	0.015 (0.008)	0.022** (0.004)	0.015 (0.011)	0.022 (0.012)	0.033** (0.006)
Quintile 5	0.010 (0.008)	0.022* (0.008)	0.029** (0.004)	0.024* (0.012)	0.035** (0.012)	0.047** (0.006)
Teacher Sample	Praxis II sub-sample		Composite z sample	Praxis II sub-sample		Composite z sample
R^2	0.66	0.66	0.68	0.7	0.7	0.71
Sample Size	173,522		1,081,142	174,589		1,087,226

**, *: Significant at 1% and 5% confidence level, respectively.

Note: Sample sizes reflect student-teacher observations in each of the teacher samples indicated. The models include the student controls described in Table 2, as well as the following teacher controls: a teacher's race/ethnicity, gender, years of teaching experience, whether a teacher is certified by the National Board for Professional Teaching Standards, license type (Continuing, Temporary, Provisional, Initial), whether a teacher received a degree from an education program approved by and located in the state of North Carolina, whether a teacher has an MA or higher degree, and selectivity of the college from which a teacher graduated.

Table 4. School and Student Fixed-Effects Specifications
(robust standard errors in parentheses)

Panel A. Full Teacher Sample						
Composite Z-score Quintiles	**Reading**			**Math**		
	1	**2**	**3**	**4**	**5**	**6**
Teacher Licensure Performance Along the Test Distribution (reference category is bottom quintile)						
Quintile 2	0.015**	0.009*	0.008**	0.024**	0.021**	0.019**
	(0.004)	(0.004)	(0.002)	(0.006)	(0.005)	(0.002)
Quintile 3	0.021**	0.010**	0.009**	0.028**	0.024**	0.022**
	(0.004)	(0.004)	(0.002)	(0.006)	(0.006)	(0.002)
Quintile 4	0.022**	0.011**	0.005*	0.032**	0.026**	0.017**
	(0.004)	(0.004)	(0.002)	(0.006)	(0.006)	(0.002)
Quintile 5	0.029**	0.013**	0.009**	0.047**	0.037**	0.030**
	(0.004)	(0.004)	(0.002)	(0.006)	(0.006)	(0.002)
Fixed Effects	None	School	Student	None	School	Student
R^2	0.68	0.68	0.93	0.71	0.71	0.94
Sample Size	1,067,235	1,067,235	1,067,235	1,073,172	1,073,172	1,073,172
Panel B. Novice Teacher Sample						
Composite Z-score Quintiles	**Reading**			**Math**		
	1	**2**	**3**	**4**	**5**	**6**
Teacher Licensure Performance Along the Test Distribution (reference category is bottom quintile)						
Quintile 2	0.008	0.023	0.016	0.028	0.053**	0.003
	(0.015)	(0.014)	(0.036)	(0.021)	(0.018)	(0.033)
Quintile 3	0.021	0.039**	0.005	0.050*	0.077**	0.006
	(0.015)	(0.014)	(0.033)	(0.020)	(0.018)	(0.031)
Quintile 4	0.022	0.036*	0.003	0.046*	0.071**	0.006
	(0.015)	(0.014)	(0.035)	(0.020)	(0.018)	(0.035)
Quintile 5	0.027	0.035*	-0.008	0.069**	0.089**	0.001
	(0.015)	(0.015)	(0.038)	(0.020)	(0.019)	(0.035)
Fixed Effects	None	School	Student	None	School	Student
R^2	0.68	0.69	0.99	0.7	0.73	0.99
Sample Size	75,554	75,554	75,554	76,019	76,019	76,019

**, *: Significant at 1% and 5% confidence level, respectively.

Note: Sample size reflect student-teacher observations. Models include the same set of controls as those specified in Table 3, with the exception of the student fixed-effects models (columns 3 and 6), which do not include a lagged test score nor time-invariant student characteristics.

Figure 1. Observed Relationship Between Licensure Test Performance and Effectiveness