Am Educ Res J OnlineFirst, published on December 2, 2009 as doi:10.3102/0002831209348970

American Educational Research Journal
Month XXXX, Vol. XX, No. X, pp. 1–34
DOI: 10.3102/0002831209348970
© 2009 AERA. http://aerj.aera.net

Race, Gender, and Teacher Testing: How Informative a Tool Is Teacher Licensure Testing?

Dan Goldhaber
Michael Hansen
University of Washington

Virtually all states require teachers to undergo licensure testing before participation in the public school labor market. This article analyzes the information these tests provide about teacher effectiveness. The authors find that licensure tests have different predicative validity for student achievement by teacher race. They also find that student achievement is impacted by the race/ethnicity match between teachers and their students, with Black students significantly benefitting from being matched with a Black teacher. As a consequence of these matching effects, the uniform application of licensure standards is likely to have differential impacts on the achievement of White and minority students.

KEYWORDS: licensure testing, teacher quality, student achievement, labor market discrimination

The role of teachers in education has long been identified as the most significant of all school factors that affect student learning, starting with the Coleman report, *Equality of Educational Opportunity* (Coleman, 1966). More recent research analyzing the relationship between teacher effectiveness and student achievement at the micro-level has confirmed these findings and further suggests that effectiveness varies considerably among teachers (Aaronson, Barrow, & Sander, 2007; Hanushek, 1992;

DAN GOLDHABER is a professor at University of Washington, Center on Reinventing Public Education, 2101 North 34th Street, Suite 195, Seattle, WA 98103; e-mail: *dgoldhab@u.washington.edu*. His work focuses on issues of educational productivity and reform at the K-12 level, and the relationship between teacher labor markets and teacher quality.

MICHAEL HANSEN is a research associate in the Education Policy Center at the Urban Institute, Washington, D.C.; e-mail: *mhansen@urban.org*. His most current research projects investigate teacher contracts, behavioral responses to incentives, and value-added estimates of performance.

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004).[1] Given the importance of teachers' input, it is understandable that policymakers attempt to guarantee a minimum level of quality through a licensure system (also commonly referred to as *teacher certification*) that delineates a set of requirements that individuals must meet in order to be eligible to teach in public schools. All states, for example, require teachers to hold a bachelor's degree and have some training in pedagogy in order to be licensed to teach, and most states require teachers to have training in the subject they teach and some kind of student-teaching experience. Teachers typically also have to pass state-mandated licensure tests before they can work in the classroom (Rotherham & Mead, 2004), and these tests are the subject of our analysis here.

In spite of the popularity of these testing policies (all but three states require teachers to pass some kind of licensure test) and the increased emphasis of testing teachers under the No Child Left Behind Act of 2001 (U.S. Department of Education, 2006), surprisingly little empirical evidence is available about the predictive validity of teachers' performance on these tests as an indicator of classroom effectiveness. Both theoretical (Stigler, 1971) and empirical work (Kleiner & Kudrle, 2000) in contexts outside of education conclude that licensure in general is not a guarantee of service quality, and, as we describe below, there is relatively little empirical work linking teacher licensure test scores to student achievement.[2]

Furthermore, no study that we are aware of has explicitly analyzed the relationship between teacher licensure test performance for specific populations of teachers. This is an important omission since minority teachers (Black teachers in particular) tend to perform substantially less well on licensure tests than do White teachers; thus, these tests have a disparate impact on who is eligible to teach. This means that licensure policies, to at least some extent, conflict with the recruitment of minorities into teaching—a long-standing policy goal, particularly in districts with large percentages of minority students (Dometrius & Sigelman, 1988; Kirby, Naftel, & Berends, 1999).

The disparate impact of teacher testing has in fact been the focus of court challenges in Alabama, California, South Carolina, and Texas, where plaintiffs have sued school districts and states based on claims of discriminatory testing practices.[3] Only in Alabama did the court rule in favor of the teachers, but these court challenges clearly suggest that state education leaders need to take the impact that these tests have on minority candidates seriously and offer evidence supporting a valid relationship between teacher testing and student outcomes.[4] Clearly, teacher testing is relevant to arguments about the diversity of the teacher workforce, but testing policies may also influence the learning of minority students; the argument for recruiting a diverse teacher workforce rests, in part, on positive gains in

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

student achievement due to matching students with teachers of a similar racial or ethnic background (e.g., Dee, 2004, 2005).

In this article, we focus on two questions regarding licensure testing policies. First, we investigate whether the information conveyed through these tests is consistent across various teacher demographic groups. The tests we analyze here are the Praxis II tests, which are commonly prescribed by licensure policies in many states (we provide more information on these tests in the data section below). Licensure tests may provide information about teachers' ability through one of two ways: screening or signaling. The test's screening function gauges how accurately the state's minimum score prevents incompetent teachers from entering the workforce while permitting more able teachers to participate. Its signaling function assesses whether a teacher's actual score on the test offers a valid prediction of student learning. Most prior studies implicitly assume that the information provided through these avenues is uniform across teachers, which may not necessarily be the case. In the results presented here, we find that licensure tests' primary function—as a screen—appears to be uniformly applied across demographic differences among public teachers but is only marginally effective at identifying teacher quality. When looking at the signaling function of the tests, however, our results show that the two tests we investigate seem to function quite differently for specific demographic subgroups in the teacher labor market: Black and male teachers.

Second, we analyze the interaction of teacher and student demographics to determine whether diversity in the teacher labor market appears to influence the distribution of student learning gains. In this study, we find significant and positive matching effects due to pairing students with same-race teachers, indicating that workforce diversity may have some academic benefit for minority students. In summary, our findings show that licensure tests provide only limited evidence of teacher effectiveness, which varies across demographic groups, and that enforcing strict cutoffs has the potential to both adversely affect minority student outcomes and decrease workforce diversity.

We begin by addressing prior research on licensure testing, student achievement, and the demographics of the public teacher workforce. Following this, we discuss our methodological strategy and the data from North Carolina used in this study. Next, we present the results of our empirical tests and perform a series of checks to ensure our findings are robust. We conclude with a summary of our findings and a discussion of the implications of licensure testing policies on the workforce and student outcomes.

Testing, Achievement, and Demographics

A teacher who instructs students operates along many different dimensions; for instance, the subject matter, the methods used, and the actual

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

learning of students all contribute to the quality of that teacher's input in the learning process (Fenstermacher & Richardson, 2005). For policy purposes, though, student learning as demonstrated through standardized tests is the most objective measure of a teacher's input that can be uniformly applied to a given pool of teachers and thus is the metric for teacher effectiveness that we use throughout our analysis here (we use teacher *quality* and *effectiveness* interchangeably throughout). This measure is consistent with Fenstermacher and Richardson's definition of successful teaching. A foolproof method of objectively identifying teacher quality, however, is elusive—research has linked teaching experience, National Board certification, and college selectivity with teacher effectiveness (e.g., Clotfelter, Ladd, & Vigdor, 2007), but variation in these observable traits explains only a small degree of the total observed variance in teacher quality (Aaronson et al., 2007).

Teacher-testing policies are intended to ensure some measure of quality control over the teacher workforce without directly observing classroom performance; but, like the other observable teacher attributes mentioned above, variation in test scores explains little of the variation in teacher quality. Three recent studies examining this relationship find a consistent, albeit small, relationship between teacher performance on licensure exams (the Praxis II tests used in North Carolina) and student achievement.[5] Clotfelter et al. (2006) find that a one–standard deviation increase in teacher performance is predicted to increase fifth-grade students' achievement by 0.01 to 0.02 of a standard deviation. Similarly, Goldhaber (2007) finds that higher licensure test performance is generally correlated with student learning gains but that the strength of the relationship depends not only on which subject the students are tested in (math or reading) but also on the specific teacher test in question (as states commonly require more than one test). Furthermore, this study concludes that the use of tests as a condition of employment eligibility likely produces many false positives and false negatives. Clotfelter et al. (2007) perform analyses in another study on the relationship between student outcomes and teacher test performance and show that the relationship is nonlinear across the distribution of teacher test scores. We expand on these studies by analyzing the consistency of the information that these tests provide for various teacher populations that make up the labor market.

Another area of debate surrounding teacher licensure focuses on how testing requirements affect the supply of teachers and, of particular concern in this study, the supply of minority teachers. At the heart of any licensure test requirement is the exclusion of individuals from the pool of potential employees: Individuals not passing the test are believed to be of unacceptably low quality and are thus deemed ineligible to teach.[6] Insomuch as diversity in the public teacher workforce is socially desirable, test standards that disproportionately screen out minority teacher candidates conflict with efforts to diversify the workforce. Gitomer and Latham (2000) analyze these

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

disproportionate failure rates on the Praxis exam and their impact on teacher supply and conclude that raising cut scores on these exams will likely result in a decrease in diversity if not somehow amended through policy intervention. A more recent study finds that the overall academic quality of teacher candidates has improved over the last decade; however, differences between White and Black pass rates on the Praxis exams have also increased (Gitomer, 2007).

A separate strand of literature raises the concern that the signaling function of licensure tests may operate differently across teacher demographics. There is, for instance, some evidence of test prediction bias (when a test differentially predicts performance, such as college or job performance) for individuals from different races/ethnicities. The SAT test, for example, tends to overpredict the academic performance for Black college students (i.e., Black students do not perform as well in college as White students, given the same incoming SAT score; Vars & Bowen, 1998).[7] In the teacher labor market, the National Teacher Examination, the predecessor to the Praxis II tests commonly used for licensure testing today, was criticized for having a cultural bias in test items that disproportionately screened Black and other minority teachers from the teacher candidate pool (Irvine, 1988; Medley & Quirk, 1974).

There are also some reasons to believe that licensure tests might differentially predict classroom performance based on the composition of students being taught, since the pedagogical or content knowledge required to make teachers successful is likely to vary depending on the background or achievement level of students. Many studies argue for a match between the race/ethnicity of teachers and students because ethnic teachers can provide a role model for, and are more likely to have high expectations of, minority students in high-needs environments. Moreover, there are important cultural differences in instructional strategies and interpretation of the behavior of minority students (Dilworth, 1986; Farrell, 1990; Fordham & Ogbu, 1986; Irvine, 1992; King, 1993; Ogbu, 1974, 1978, 1992; Villegas & Clewell, 1998; Zapata, 1988; Zimpher & Ashburn, 1992). Dee (2004) provides empirical evidence for these claims by using data from an experiment where students were randomly matched to classrooms (and therefore, he argues, teachers) and finds positive benefits for what he calls *role model matching*, when students and teachers of the same race are paired together (we use the more general label *matching effect* in this study, as this learning benefit may be transmitted to students through cognitive as well as social means). Dee estimates that students paired with a same-race teacher increase performance on standardized tests by 2 to 3 percentile points and finds that the effect accumulates over time.

The empirical evidence on an analogous matching effect is not nearly as conclusive when matching is done along gender lines. Ehrenberg, Goldhaber, and Brewer (1995) find that student gender affects teachers' subjective evaluations of students, but they find no evidence that it affects students' test scores. At the college level, D. S. Rothstein (1995) finds a positive

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

relationship between the level of postgraduate educational attainment in females and the percentage of female faculty members in the majoring departments of their undergraduate institutions. Lavy (2004) uses a natural experiment in Israel to find a gender stereotype against boys, likely resulting from teachers' (not students') behavior; he estimates that the magnitude of the bias against boys amounts to standardized test scores 0.05 to 0.18 standard deviations lower than otherwise expected, depending on the test subject. Dee (2005, 2007) finds that student gender does play a significant role in teachers' perceptions of student performance, student test scores, and student engagement in academic subjects; he estimates that same-gender matching accounts for approximately 0.05 standard deviations in student performance on literature, math, and science tests. Holmlund and Sund (2008) analyze student performance in Sweden and find statistically significant performance gains of 0.05 standard deviations among female students assigned to female teachers in natural sciences, but in contrast to Dee's results, they fail to find a matching effect across all subjects.

While the matching findings show that student achievement gains (particularly among minority students) can be leveraged to encourage the recruitment of more minorities into teaching, this policy goal appears to be in conflict with what some see as the need to upgrade the skills of teachers of minority students. Ferguson (1998), for instance, recognizes these conflicting policy goals and strongly argues for higher quality standards rather than workforce diversity, as the better-qualified pool of teachers, in his view, stand to disproportionately benefit disadvantaged students. In this article, we revisit this issue and present some evidence on the correlation between teacher test performance and student outcomes, particularly analyzed across the demographics of the teacher workforce and the student body. We hope the evidence presented here will help to inform policymakers about the efficacy and utility of teacher-testing policies.

Methodological Approach

States use licensure tests primarily as a screening device: Teachers must pass the test to be eligible to teach.[8] But from a policy perspective, we might also be interested in the information they provide as a signal of teacher classroom effectiveness. In the empirical models that follow, we distinguish between the screening function and the signaling function of these tests.

We estimate the effectiveness of the tests as a screen by analyzing their ability to successfully identify effective teachers with the following model:

$$A_{i,j,t} = \alpha A_{i,t-1} + \beta_1 \text{STUDENT}_{i,j,t} + \beta_2 \text{CLASS}_{j,t} + \delta \text{PASS}_{j,t}, \quad (1)$$

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

where the dependent variable here is the achievement of student *i*, in classroom *j*, in year *t*. The model includes controls for achievement in a prior year, $A_{i,t-1}$; a vector of individual student characteristics, STUDENT; a vector of classroom variables, CLASS; and an indicator for whether the student's teacher passed or failed the licensure standard, PASS. Note that we do not include other teacher-level attributes, because the licensure policy screens on the basis of test performance only, without considering any other alternative indicators of teacher ability. The main parameter of interest here is the estimate of δ, which is identified by the comparison of students with teachers who pass the licensure test standard to those whose teachers do not pass, holding constant the other variables in the model. Larger estimated values of δ suggest greater value in the screen, as the screen prevents less-effective teachers from participating in the labor market. Estimates that are not significantly different from zero suggest little to no screening value, since passing (or failing) the standard is not related to how effective (or ineffective) a teacher is in the classroom.

We wish to know whether the quality of the screen is invariant to the group of teachers (either racial or gender group) being measured. Most critiques of screening policies based on discriminatory grounds suggest that policies may be effective in screening out most potentially ineffective teachers but do not function as well for minority groups (e.g., Dilworth, 1986). Our first test addresses these differential screening values to detect how the impact varies across demographic groups in shaping the labor force. For the purpose of comparison, we isolate teachers according to race or gender, estimate the model above, and test for differences from the more-restrictive pooled model (which constrains all teachers to share the same estimate on the pass-fail indicator, without regard to demographic differences). Understanding how the efficacy changes across race or gender provides important information that can inform policy decisions about extending licensure contingent upon test performance.

Next, the Educational Testing Service, the organization that developed and administers the Praxis II tests that we analyze in our study, claims that its commonly used teacher tests may not have great predictive power for teachers who score above a given state's cut score (because the purpose of the tests focuses on minimum competencies and less on differentiating between individuals at the upper end of the distribution).[9] But there is no national cut score, and the fact that scores vary considerably between states (Mitchell, Robinson, Plake, & Knowles, 2001) warrants a closer look at the predictive power of licensure tests along the entire performance distribution. The variation in scores is certainly relevant for policy, as a local school district, at the point of hire, may use a teacher's score (as opposed to the simple pass or fail status) as a signal of teacher quality.

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

To explore the predictive validity of the tests across teacher demographics we employ the following model:

$$A_{i,j,t} = \alpha A_{i,t-1} + \beta_1 \text{STUDENT}_{i,j,t} + \beta_2 \text{CLASS}_{j,t} + \beta_3 \text{TEACHER}_{j,t} + \gamma \text{SCORE}_{j,t}. \quad (2)$$

This is a simple variant of Equation 1, but here, instead of controlling for whether a student's teacher passes the current standard, we control for the teacher's performance on the exam. In this model, we also include a vector of teacher characteristics (TEACHER) since local districts, when hiring teachers, have the benefit of assessing teacher test scores in the context of additional information about them (for instance, a teacher's degree level or college selectivity). As above, our primary interest is in estimating this model separately for each identifiable teacher subgroup (by race or gender). Here, we are concerned with the estimates on γ for each group: Positive estimates of γ indicate a positive relationship between test performance and resulting student achievement, while estimates that are not different from zero would indicate that test performance provides no information about a teacher's effectiveness in the classroom. Again, if these tests were consistently calibrated as a measure of teacher effectiveness across demographics, we would expect the marginal increase in student achievement resulting from higher test performance to be reasonably constant across teacher demographic traits.

The two models above attempt to assess the consistency of the information conveyed through teacher candidates' performance on the Praxis II tests, as measured in student outcomes. The conflicting policy goal, however, regards the diversity of the workforce. Diversity among public school teachers has important social value, certainly, in encouraging tolerance, providing minority role models, and facilitating other socially desirable ends. Our focus, however, is to analyze the effect of diversity on the academic achievement of students. If diversity in fact has a positive effect on achievement, policymakers are potentially faced with a tradeoff between the academic benefits of diversity (either through same-race or same-sex matching) and the academic benefits of a better-qualified labor force (through higher testing standards). On the other hand, if we detect no (or negative) student gains due to diversity, then the social benefits of diversity alone must be weighed against the cost of lower-achieving students.

To quantify the academic effect of a more-diverse workforce, we assess how teachers with different demographic characteristics affect students of various backgrounds, as matching relationships assume a positive benefit to same-race or same-gender matches between teachers and students. For this, we repeat a model similar to the signaling model (Equation 2 above), simply isolated on student race and gender characteristics. For instance, we analyze the effects of all teachers on Black students, allowing Black and White teachers to have different effects; this allows for different effects along demographic lines in both the students and the teachers. The only variation in the estimated

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

equation is the presence of an intercept and interacted slope coefficient for Black or male teachers (depending on whether we are making a cross-race or cross-gender comparison). The intercept estimates an average impact of the matching relationship, and the slope coefficient allows us to estimate how this relationship may change over the test score distribution. In these models, the estimated coefficients are identified based on variation in teacher test scores among those teaching a particular student group, and it reveals the impact of specific student-teacher interactions along either same-race or same-gender matching. Having established the methodological approach, we now describe the data that we use.

Data

The data used in this study are derived from North Carolina Department of Public Instruction administrative records, as maintained by the North Carolina Education Research Data Center. The records cover the universe of teachers and students in the state over an 11-year period (spanning school years 1994–1995 though 2004–2005). The data allow us to link students and teachers (at the elementary level) and track them over time.

The student achievement measures come from state-mandated, standardized end-of-grade reading and math tests as prescribed in the North Carolina Standard Course of Study (North Carolina Department of Public Instruction, 2007b). These criterion-referenced tests are vertically aligned to enable the North Carolina Department of Public Instruction's Accountability Department to assess performance, growth, goals, and ratings for all schools in compliance with the state's ABC education reform program. Students' test scores are transformed to have a mean of zero and a standard deviation of one for each grade and year to allow for comparisons across multiple grades and years. The data also include background information on individual students, such as gender, race and ethnicity, parental education, disability, and eligibility for (not necessarily receipt of) the federal free and reduced-price meals service.

The data do not permit a direct link from students to their teachers but rather link students to the teacher who monitored their exam. In most cases, this monitor teacher is the students' classroom teacher, although we cannot directly externally validate this assertion with the data. There are, however, a number of ways that we can assess the internal validity of this assertion through multiple checks on the data that are consistent with having the monitor teacher as a regular teacher during the school year. For instance, for our analysis we use only monitoring teachers who are documented in their personnel files as teaching a self-contained class (where one teacher spends the majority of each instructional day with the same group of students) of the same grade level.[10] We indirectly check on the validity of the teacher-student match by comparing our findings to those from other research that includes

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

at least some common variables. In fact, our findings on the returns to teacher experience are very consistent with those from studies that do directly link students and teachers (e.g., Rivkin et al., 2005; Rockoff, 2004), and the magnitude of estimates of the value of having an National Board for Professional Teaching Standards–certified teacher are similar to estimates reported by Sanders, Ashton, and Wright (2005), who analyze a subset of districts from North Carolina where they were able to directly link teachers and students.[11] It is highly unlikely that there would be many cases where we have inappropriately classified a monitor teacher as a student's teacher: Were this to happen, our estimates on these variables would suffer an attenuation bias (relative to those in the prior literature) resulting from any mismatch.

We are also restricted to students in Grades 4 through 6 who have a valid math and/or reading pre- and posttest score (e.g., the end-of-year fourth-grade math score is used as the pretest when a student's end-of-year fifth-grade math score is the dependent variable).[12] We also eliminate classrooms with fewer than 3 or more than 28 students (corresponding to the 1st and 99th percentiles in our data), as these classrooms on the extremes of the distribution may not represent the most common classroom environment that we attempt to isolate here.[13] Our primary motivation in applying these restrictions is to identify a group of students who, for the purposes of this analysis, are highly likely to be matched to their teachers of math and reading in a typical classroom setting.

The teacher data include information on teachers' degrees and experience levels, licensure status, teaching assignment, and the college from which the teacher graduated, in addition to teachers' performance on one or more licensure exams. Educational Testing Service's Praxis II exams, a sequence of tests that focus on specific subject matter knowledge and/or pedagogical preparation, are the tests prescribed by North Carolina's licensure policies (for more information, see the Educational Testing Service Web site: http://www.ets.org/praxis). As of July 1997, elementary school teachers in North Carolina were required to meet a minimum performance standard on two specific Praxis II tests (Test Nos. 10011 and 20012): the Praxis II Curriculum, Instruction, and Assessment test (CIA) and the Praxis II Content Area Exercises test (CAE).[14] Teachers entering the workforce laterally after July 1997, however, could meet the testing requirement with acceptable previous scores on either the National Teacher Examination (Educational Testing Service's predecessor to the Praxis II tests) or the Graduate Record Exam in lieu of the Praxis II tests.

Because teachers' scores on these Praxis II tests are key to our study, our student sample is again restricted to include only those whose teachers have valid Praxis II test scores on both the CIA and CAE exams. When we include scores in our estimated models, all Praxis II scores are normalized (to have a mean of zero and a standard deviation of one) relative to all other teachers taking the exams in the same year. In Table 1, we document how our data

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

Table 1
Selection of Data Sample From Universe of Unrestricted Data

Panel A: Case Count of Data Selection by Each Restriction	Observations	Unique Students	Unique Classes	Unique Teachers
Universe of data	4,267,153	1,679,740	201,859	65,688
Self-contained class	2,426,403	1,242,464	113,345	31,356
Valid pretest scores	1,268,296	779,117	76,848	24,638
Class size restriction	1,239,461	771,225	62,574	19,604
Not missing values	1,223,524	761,429	62,435	19,513
Valid Praxis scores	193,725	174,828	9,760	4,051
Panel B: Descriptive Characteristics of Students (Unweighted Observations)	**Universe of Data**	**Linked Sample**	**Restricted Sample**	
Percentage female	.487	.495	.496	
	(.500)	(.500)	(.500)	
Percentage White	.605	.629	.588	
	(.489)	(.483)	(.492)	
Percentage Black	.301	.290	.310	
	(.459)	(.454)	(.463)	
Percentage eligible for free or reduced-price meals service	.468	.443	.468	
	(.499)	(.497)	(.499)	
Percentage with parent(s) holding bachelor's degree or higher	.159	.169	.216	
	(.366)	(.375)	(.412)	
Total	4,267,153	1,223,524	193,725	
Panel C: Descriptive Characteristics of Teachers (Unweighted Unique Classrooms)	**Universe of Data**	**Linked Sample**	**Restricted Sample**	
Percentage female	.907	.915	.864	
	(.290)	(.279)	(.343)	
Percentage White	.834	.838	.857	
	(.373)	(.369)	(.350)	
Percentage Black	.153	.150	.127	
	(.360)	(.357)	(.333)	
Mean experience	12.825	12.778	3.176	
	(9.775)	(9.911)	(4.273)	
Percentage whose highest earned degree is a bachelor's degree	.708	.720	.868	
	(.455)	(.449)	(.339)	
Total	201,859	62,435	9,760	

Note. Estimated standard deviations in parentheses.

sample was selected from the unrestricted universe of data on elementary students provided by the North Carolina Education Research Data Center.

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

As shown, the restriction on self-contained classrooms and valid pretest scores is the most restricting criterion that we impose to arrive at the teacher-student linked sample. We then compare the attributes of all student observations in the universe of data, those we are confident in linking to their classroom teachers, and those whose teachers have valid Praxis II scores (which constitute our restricted sample). We report similar descriptive statistics for the teachers in our data sample as well. Because North Carolina's licensure policy requiring teachers to pass the Praxis II exams began in 1997, most of the teachers in the restricted sample are understandably less experienced.

In 1997, North Carolina's minimum score standards on the CIA and CAE tests were 153 and 127, respectively. In 2000, the state eliminated separate test score minimums and replaced them with a minimum combined score of 313 on the two tests; however, teachers who already entered the labor force under the previous standard were grandfathered in regardless of whether they actually met the new standard. In addition, teachers may teach in a North Carolina school without meeting the Praxis II requirement, with a temporary license that is valid for 1 year, but must then achieve an acceptable score on the Praxis II in order to continue teaching (North Carolina Department of Public Instruction, 2007a). Taken together, we have two sources of variation that provide teachers' scores below the level required to pass the current standard, which provide the basis for identifying the screening value of the exams in addition to providing a greater range for determining their signal value.

In total, our data set includes 4,051 unique teachers (9,760 teacher observations) with 174,828 unique students (193,725 student-teacher observations). Table 2 presents descriptive statistics for all teachers in our data set on various teacher, student, and classroom characteristics that are used in the analyses below. The teachers are grouped according to race and gender, for the purpose of comparing along these lines.[15] The means reported for time-varying variables (such as school-level minority and free lunch measures) are first averaged over each teacher's tenure in the North Carolina system; these values are then averaged across all teachers to arrive at the means reported here.

It is worth highlighting the significant differences in the mean performance of the Praxis II exams between races.[16] Given Educational Testing Service's documentation of a persistent performance gap between White and Black teacher candidates (Gitomer, 2007), we are not surprised to see a corroborating gap between the groups on these tests—almost 0.80 standard deviations on the CIA test and over 0.50 standard deviations on the CAE test. A moderate gap between male and female performance is also evident in the data but not to the extent of the gap between White and Black teachers. We also report the number of unique teachers in each group who either currently fail North Carolina's licensure standard (because they

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

Table 2
Descriptive Statistics by Teacher Demographic Group

Test Performance and Effectiveness Estimates	White	Black	Female	Male
Praxis CIA performance (normalized)	.390	−.399	.294	.206
	(.548)	(.818)	(.649)	(.635)
Praxis CAE performance (normalized)	.292	−.241	.249	.011
	(.751)	(.842)	(.784)	(.743)
Student learning in reading	−.010	.082	.004	−.032
	(0.966)	(1.178)	(1.006)	(0.954)
Student learning in math	−.004	.025	−.004	.025
	(0.986)	(1.066)	(0.982)	(1.121)
Professional Characteristics				
Years of experience	2.530	3.536	2.698	2.386
	(3.787)	(4.664)	(3.948)	(3.620)
Percentage holding master's degree	.125	.122	.123	.130
	(.319)	(.313)	(.318)	(.322)
Percentage fully licensed	.590	.564	.588	.571
	(.462)	(.465)	(.464)	(.454)
Classroom Characteristics				
Percentage of minority students	.409	.740	.454	.466
	(.272)	(.249)	(.292)	(.285)
Percentage of students eligible for free or reduced-price meals service	.464	.660	.491	.494
	(.238)	(.254)	(.252)	(.237)
Average class size	19.457	18.112	19.181	19.952
	(4.597)	(5.389)	(4.834)	(4.039)
Total	3,460	522	3,560	491
Unique failing teachers	235	203	367	80

Note. The results are for a subsample of teachers. The weighted mean of the standardized licensure score across teachers in this sample is not zero. It is above zero, because the teachers in this subsample outperformed the teachers in the state on the licensure test. Estimated standard deviations in parentheses. CIA = Praxis II Curriculum, Instruction, and Assessment test; CAE = Praxis II Content Area Exercises test.

were grandfathered into the current standard or are teaching on a temporary license) or previously taught for 1 or more years with failing scores before successfully passing.[17] We see here disproportionate failing rates between races: Unique Black teachers constitute 13% of our sample but account for 45% of those who do not qualify in terms of achieving the cut score standards. By contrast, professional credentials, including years of experience, graduate degree attainment, and full licensure, are roughly equivalent across all teachers in our study.[18] Inspection of the classroom characteristics of these teachers also reveals that Black teachers tend to work in more-disadvantaged teaching assignments compared with White teachers: Average measures of minority student percentages and students

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

eligible for free and reduced-price lunch show remarkable differences between these two groups. Similar comparisons between male and female teachers show their teaching assignments to be essentially the same.

Table 2 also reports estimates of the mean for each teacher category of the value-added contribution of teachers toward student achievement. These teacher effects are estimated based on a fixed effects regression of student achievement that controls for student characteristics, including the student's pretest score, and a teacher fixed effect.[19] Comparison of these estimates across racial differences shows that Black teachers' students, on average, outperformed White teachers' students in both reading and math, holding observable student attributes constant (this difference is significant in reading). This relationship holds in spite of Black teachers' poorer average performance on licensure tests (none of these performance differentials in the classroom, however, are as large as that of the performance differential on the test). Inspecting this difference along gender lines, we see that male teachers score lower (on average) in reading and higher in math compared to female teachers (though these differences are not significant).

The discrepancy between teacher test performance and effectiveness is more obvious when presented visually, as in Figure 1, which graphs the mean confidence ellipses of combined performance on the Praxis tests on the *x*-axis with estimated teacher effectiveness in math on the *y*-axis. These ellipses represent the limits of a two-way .95 confidence interval jointly drawing Praxis II performance and estimated teacher effectiveness in math. The graphs demonstrate three important attributes of this relationship. First, the confidence ellipses are more circular than elliptical, suggesting a weak correlation between these measures. Second, the vertical range (representing classroom effectiveness) is virtually the same for the all groups, while the horizontal spread (representing test performance) clearly places the White and female teachers ahead of Black and male teachers. Third, because of this discrepancy, virtually any cut score imposed on the labor market disproportionately screens out Black and male candidates (as demonstrated by the reference line for the North Carolina standard). This provides at least cursory evidence that licensure tests may not accurately predict teacher effectiveness across teacher types. We explore this issue in more detail in the models described below.

Empirical Results

Table 3 reports the average differences in student achievement (in reading or math) between teachers who pass the current North Carolina standard and those who fail, holding all student and classroom characteristics equal. Teacher characteristics are not held equal in this specification, as we wish to estimate this difference without regard to other teacher qualifications, just as testing policies are blind to other teacher attributes.

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

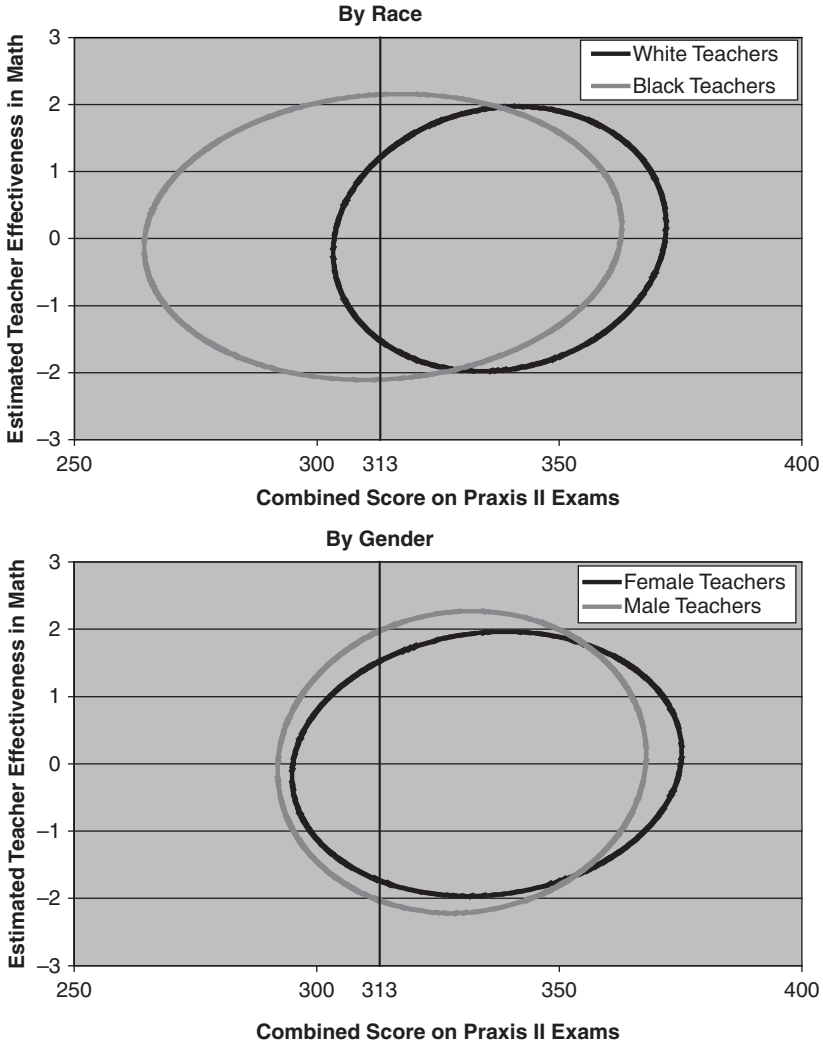

Figure 1. **Confidence ellipses of relationship between Praxis performance and student achievement.**

The estimates are reported first from a restrictive pooled model and then broken out by different race or gender. Given that the majority of unique teachers in our data set are female (88%), White (85%), or both (75%), we are not surprised to see the estimates for passing the screen are roughly equal between these subgroups and the pooled model. Our estimates are

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

Table 3
Differential Screening Values for Passing Licensure Standard

Panel A: Screening Value on Reading Scores	Pooled	White	Black	Female	Male
Passing current testing standard	.017†	.021†	.008	.018†	.015
	(.010)	(.012)	(.016)	(.010)	(.026)
Observations	192,354	166,179	23,124	165,641	26,713
R^2	.66	.66	.64	.66	.65
Panel B: Screening Value on Math Scores					
Passing current testing standard	.049**	.052**	.041*	.058**	.012
	(.013)	(.016)	(.021)	(.014)	(.031)
Observations	193,505	167,137	23,296	166,641	26,864
R^2	.70	.71	.68	.71	.70
Teachers failing standard	447	235	203	367	80

Note. Robust standard errors in parentheses clustered at the teacher level. Estimates from least squares regressions with the following control variables: indicators on matching relationship, student's race and gender, grade, learning disabilities, free lunch eligibility, limited-English proficiency status, parents' education level, class size, year, and the percentage of minority students in the classroom.
†$p \leq .10$. *$p \leq .05$. **$p \leq .01$.

consistent with previous findings between licensure performance and student achievement (Goldhaber, 2007); specifically, there is no significant estimate on the screening function when measuring the gains to student performance in reading, but passing is significantly associated with an increase of approximately 0.05 standard deviations in student achievement in math. By isolating each teacher subgroup and estimating the same model, we find that the value of the screen comes primarily from the dominant teacher demographic. The value of the screen has lower point estimates for the other demographic groups of interest—Black teachers and male teachers—but in neither case can we say that the estimates are significantly different from those estimated in the White or female regressions.[20]

Our results for these first tests suggest that the estimated difference in returns for passing the licensure standard on reading achievement is not significantly different from zero at any level of analysis—either pooled or taken by each teacher subgroup. When gauging the returns in math achievement, we have a positive and significant screening value to the licensure standard among all teachers (with the exception of male teachers, where the effect is positive but not significant); furthermore, an *F* test for the joint equality of passing the test fails to reject the hypothesis that the magnitude of passing the licensure standard is constant across demographic subgroups. In light

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

Table 4
Differential Signal Values of Praxis Exams

Panel A: Comparison of Signal Value on Student Performance in Reading	Pooled	White	Black	Female	Male
Praxis CAE performance	.003	.002	.001	.002	.012
	(.003)	(.003)	(.011)	(.003)	(.010)
Praxis CIA performance	.008†	.013**	−.013	.008†	.005
	(.004)	(.005)	(.011)	(.005)	(.012)
Observations	192,354	166,179	23,124	165,641	26,713
R^2	.66	.66	.64	.66	.66
Panel B: Comparison of Signal Value on Student Performance in Math					
Praxis CAE performance	.009†	.005	.024†	.006	.019
	(.005)	(.005)	(.014)	(.005)	(.012)
Praxis CIA performance	.011†	.018*	−.006	.016*	−.010
	(.006)	(.007)	(.014)	(.006)	(.019)
Observations	193,505	167,137	23,296	166,641	26,864
R^2	.71	.71	.68	.71	.70

Note. Robust standard errors in parentheses clustered at the teacher level. Estimates from least squares regressions with the following control variables: indicators on matching relationship, student's race and gender, grade, learning disabilities, free lunch eligibility, limited-English proficiency status, parents' education level, teacher's race and gender, college selectivity, license, National Board certification, educational attainment, experience level, graduation from a North Carolina–approved education program, class size, year, and the percentage of minority students in the classroom. CIA = Praxis II Curriculum, Instruction, and Assessment test; CAE = Praxis II Content Area Exercises test.
†$p \leq .10$. *$p \leq .05$. **$p \leq .01$.

of these estimates, we find no evidence suggesting that the screening function of the test is discriminatory on the basis of its utility in retaining competent teachers while screening out those of low quality. We wish to note, however, that the screen is equally ineffective for all teachers on the basis of reading achievement.

Next we test to see whether Praxis performance (not just a pass-fail indicator) is equally informative as a signal of teacher effectiveness across all teachers. Our results from these tests are summarized in Table 4.

As with the findings presented in Table 3, we find that much of the signal value of the tests in the pooled model is attributable to the dominant groups in the teacher pool. Student achievement in both reading and math is significantly correlated (at the .10 alpha level) with performance on the CIA Praxis exams in the pooled model, and student performance in

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

math is additionally correlated with CAE performance. Furthermore, in both reading and math, the point estimates on CIA performance are higher than those of performance on the CAE portion.

Once we isolate the effects of each demographic group, however, we observe some divergence in the signal value for each of the subgroups, particularly on the CIA test. For instance, CIA performance is significantly positive for student achievement in math for the female teacher group and in both reading and math in the White teacher group. Now, regarding male teachers in math and Black teachers in both reading and math, the same performance on the CIA exam (while not significant) was estimated to be negative. Our tests do not have enough power to detect significant differences in the estimates on male and female teachers; however, we do reject the hypothesis of the equality of the estimates between Black and White CIA performance (at the .10 level) in both the reading and math specifications of the model.

Performance on the CAE portion of the exam was significant in the pooled model when using student achievement in math as the dependent variable, but when we break these into demographic groups, we find that performance is significant only in the case of Black teachers. In fact, Black performance in math student achievement was the only significant finding for any subgroup on the CAE portion (though the estimates for male teachers in both subjects was large in magnitude but not statistically significant).

While only a portion of the estimates for each subgroup is significant, the patterns we see in the estimates suggest that the two different Praxis tests do not have the same predictive validity in signaling a teacher's classroom effectiveness. In particular, our results suggest that CIA test performance is a reasonably predictive signal of quality for White teachers but not informative for Black teachers. Conversely, teacher performance on the CAE test is virtually independent of student achievement for White teachers but significantly related to student performance in math for Black teachers. The finding that the two tests operate quite differently for different racial groups is not surprising when we consider that the test format for the two is quite distinct—the multiple-choice CIA test seems to be a better predictor among White and female teachers and the essay-response CAE test provides a better gauge for performance among Black teachers. Thus, we cannot make any predictions about a teacher's quality, given his or her performance on each teacher test, without also considering the teacher's race and gender.

Related to these results (but not reported) are three auxiliary models that we ran to better understand the signal value of the tests. First, we estimated the models above including only one teacher test at a time (instead of both simultaneously). The estimates on the various teacher tests were largely unchanged (varying only in the thousandths place) from the reported results, and the patterns of significance remained the same regardless of whether each test entered the model separately or jointly with the other.

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

Second, to allow for nonlinearities within the test performance–student achievement relationship, we coded test performance as a dummy variable indicating the quintile of test performance. While these models did reveal some additional information in the relationship, we came to similar conclusions about the signal value of the tests—specifically, the signal values were not consistent across race or gender differences. Third, as with the models looking at the screening function above, we estimated intermediary models between the restrictive pooled models and the unrestrictive but mutually exclusive models. Again, tests assessing the goodness of fit across specifications led us to report the most unrestrictive models in our results, as presented here (see note 18).

As a final analysis of the relationship between test performance and student achievement across teacher demographics, we investigate the interaction between teacher test performance for these demographic groups and various demographic groups of students. We present the results in Table 5, which reports the estimates of teachers' combined performance on the two Praxis tests for Black and White teachers on student achievement. Here, students are isolated by race to test for same-race matching in Panel A and isolated by gender for same-sex matching in Panel B.

Our underlying motive for this analysis is to understand how manipulating the demographic composition of the teacher candidate pool might influence student achievement. Like the teacher candidate pool, the student body is composed of students with varying demographic characteristics, which may also respond differently to teachers of different demographics. In these models, a teacher's combined performance on the Praxis II exams (the sum of CIA and CAE exams) is normalized within year, as in previous models, and further interacted with teacher attributes (either male or Black).

Focusing first on Panel A, we find that performance on the Praxis II exams is positively and significantly correlated with White and other race student outcomes (for Black students, the estimates are positive but not significant at standard levels). The intercept for having a Black teacher is what we are using to capture potential matching effects. If there were a same-race matching effect for White students assigned to White teachers, we should see a negative coefficient in the first two columns; if there were a significant effect among Black teachers and students, we would see a positive coefficient in the third and fourth columns. The lack of a significant difference among White students on this variable suggests they are, on average, insensitive to their teacher's race. On the other hand, we see evidence that both Black and other ethnicity students are sensitive to their teacher's race—they are positively correlated with having a Black teacher in the classroom.

The slope coefficients for both Praxis variables (Praxis and Black × Praxis) also reveal how this relationship may change over the distribution of test scores. For all students we see positive (though not always significant)

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

Table 5
Estimated Same-Race and Same-Sex Matching Effects

Panel A: Same-Race Matching in Reading and Math	White Students		Black Students		Other Race Students	
	Reading	Math	Reading	Math	Reading	Math
Combined performance on Praxis	.008*	.014**	.006	.008	.021*	.016†
	(.004)	(.005)	(.005)	(.006)	(.009)	(.009)
Black	.018	.011	.043**	.035*	.062*	.047†
	(.011)	(.016)	(.012)	(.016)	(.025)	(.024)
Black × Praxis Performance	−.009	−.006	−.015	−.006	−.030	.033†
	(.009)	(.012)	(.011)	(.012)	(.018)	(.020)
Observations	112,042	112,531	58,587	59,027	9,008	9,058
R^2	.64	.68	.56	.61	.64	.71
Panel B: Same-Sex Matching in Reading and Math	Female Students		Male Students			
Combined performance on Praxis	.002	.013**	.010**	.013**		
	(−.003)	(−.004)	(−.004)	(−.005)		
Male	−.001	.006	−.007	.015		
	(.008)	(.012)	(.008)	(.012)		
Male × Praxis Performance	−.003	−.013	−.004	−.011		
	(.011)	(.012)	(.011)	(.013)		
Observations	95,740	95,985	96,614	97,520		
R^2	.66	.71	.65	.70		

Note. Robust standard errors in parentheses clustered at the teacher level. Estimates from least squares regressions with the following control variables: indicators on matching relationship, student's race and gender, grade, learning disabilities, free lunch eligibility, limited-English proficiency status, parents' education level, teacher's race and gender, college selectivity, license, National Board certification, educational attainment, experience level, graduation from a North Carolina–approved education program, class size, year, and the percentage of minority students in the classroom.
†$p \leq .10$. *$p \leq .05$. **$p \leq .01$.

student outcomes related to Praxis performance, but we estimate negative coefficients on most interacted values, roughly offsetting the signal value for Black teachers in most cases (the sum of Praxis and Black × Praxis is close to zero and not significant in every case except the last column). This suggests that the matching effect (the estimated difference between different-race teachers for the same groups of students) is not constant over the distribution but generally appears to deteriorate as one moves up the distribution of Praxis performance (due to White teachers significantly increasing in predicted effectiveness, while Black teachers' predicted effectiveness stays roughly constant).

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

These findings also show an important result that is missed in the previous tables: Same-race matching effects dwarf most any information conveyed through the licensure test signal. We wish to point out that when teaching Black students, Black teachers in the lower end of the teacher test distribution are estimated to have impacts that are approximately the same as White teachers at the upper end of the distribution.

Moving our attention to Panel B, we do not find any strong evidence of positive effects due to same-sex role modeling. The intercept terms in all cases are not significant, and the interacted coefficients again generally offset the positive relationship to Praxis performance. While none of these coefficients are significant, the general trend of the estimates suggests no advantage to either gender but an increasing advantage to female teachers as one moves up the test distribution (all things equal).

In summary, we find that evidence suggesting the uniform application of licensure standards for all teachers is likely to have differential impacts on the achievement of White and minority students. Specifically, we see that Black and other minority students appear to benefit from being matched with a Black teacher regardless of how well or poorly that teacher performed on the Praxis tests, and these positive effects due to matching with Black teachers are comparable in magnitude to having the highest-performing White teachers in the classroom. Removing the lowest of performers on the exam would necessarily remove some of the teachers that appear to be most effective for this segment of the student population.

Threats to Internal Validity

In performing these analyses, we wish to acknowledge three potential sources of bias in our estimates and describe our efforts to counter these threats to validity. These threats come from selection on unobservable teacher characteristics, nonrandom sorting of students in classrooms and teachers among schools, and nonrandom attrition from the teacher labor market.

Selection Bias

Unobserved teacher characteristics may be correlated with student achievement, and if these characteristics are not uniform across the distribution of teachers, we worry about attributing differential performance on the test to differences in these unobservable attributes. We worry that this selection problem may arise where teachers who fail the exam (on one or more occasions) may possess unobservable attributes—positive or negative—that vary systematically from teachers who passed the cutoff from the outset. This could be manifested in our data in one of two ways—first, by mitigating the class of teachers who meet the standard or, second, by a selection bias among teachers who fail the exam.

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

First, the licensure standard may in fact be effective in screening poor-quality teachers from the labor market; however, because teachers can bank scores and retake those portions of the test on which they performed poorly, the average quality of the pool of teachers who eventually meet the licensure standard may be lower than it would be if unaffected by those teachers who initially failed. If this were the case, our estimates suffer an attenuation bias, decreasing the magnitude of the true difference between passing and failing candidates. To counter this potential selection problem, we utilize available data on the Praxis test performance for each teacher at the time he or she initially enters our data set, and we classify teachers according to their initial passing status (instead of allowing it to change once higher scores on the Praxis are attained). For brevity, we do not present our estimates using the minimum score; however, our results were consistent with those already reported and suggest that this manifestation of selection bias is not an issue here.

In the second form of selection bias, we observe teachers in the classroom only if they are successful in obtaining a teaching position, and those who pass and those who fail likely have a different propensity for selection into our data set. In one case, those teachers who secure a teaching position despite having failed the licensure standard may have positive unobservable attributes that those who pass the test may not necessarily have. Conversely, teacher candidates who fail the standard may be discouraged from teaching and begin to seek employment in alternative careers, and those who continue to pursue teaching opportunities despite failing are those with the least promising prospects outside of teaching. The estimated quality of failing teacher candidates is biased upward in the first scenario (from positive unobservable attributes) and downward in the second (from marketable skills' potential correlation with teaching ability); how this affects the estimated difference is unclear.

We propose the use of a regression-discontinuity model to determine if there are any systematic differences between those who pass and those who fail in our samples. An regression-discontinuity approach combines the two models employed thus far, by including nonlinear controls on the Praxis test scores and an indicator on whether a teacher passed or failed the licensure standard. The cut score on these licensure tests varies across states, and the distribution of initial test scores (in our data, at least) appears smooth around North Carolina's cut score. Given these two points, we have no reason to believe that those teachers marginally passing the test enjoy an extra endowment of teaching ability relative to those who marginally failed the standard, after controlling for the nonlinear signal conveyed through the score. Thus, if selection bias among those in the failing group were an issue, it would manifest itself as a significant discontinuity in estimated teacher effectiveness at the cut score, where our hypothesis is that no significant difference exists between teachers on either side of the cut point. In Table 6, we report the estimated

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

Table 6
Regression Discontinuity Estimates to Detect Selection Bias

Panel A: Screening Value on Reading Scores	Pooled	White	Black	Female	Male
Estimated discontinuity	.000	−.004	−.002	−.009	.036
	(.015)	(.019)	(.031)	(.016)	(.037)
Observations	192,354	166,179	23,124	165,641	26,713
R^2	.66	.66	.64	.66	.66
Panel B: Screening Value on Math Scores					
Estimated discontinuity	.014	.021	−.038	.006	.069
	(.021)	(.026)	(.044)	(.022)	(.052)
Observations	193,505	167,137	23,296	166,641	26,864
R^2	.71	.71	.69	.71	.70
Teachers failing standard	447	235	203	367	80

Note. Robust standard errors in parentheses clustered at the teacher level. Estimates from least squares regressions with the following control variables: indicators on matching relationship, student's race and gender, grade, learning disabilities, free lunch eligibility, limited-English proficiency status, parents' education level, teacher's race and gender, college selectivity, license, National Board certification, educational attainment, experience level, graduation from a North Carolina–approved education program, class size, year, and the percentage of minority students in the classroom. Additional controls include quadratic expansion of combined Praxis score (deviation from cut score), interacted with passing status to estimate discontinuity.

discontinuities from this design. As in the prior tables, we implement this regression-discontinuity approach on the pooled data and then on each teacher demographic. None of the estimated discontinuities we present are significant; thus, we do not believe our estimates suffer from selection bias.

Nonrandom sorting

The second threat to internal validity, positive selection of teachers to schools and students to teachers, is well documented (Clotfelter et al., 2006). Since these documented sorting patterns generally result in upwardly biased teacher coefficient estimates, our estimates on licensure test performance may be upwardly biased as well. To account for this potential bias, we replicate all of our tests using school fixed effects. Using a school fixed effect approach will demean all variables (both dependent and explanatory variables) at the school level; thus, only variation off of the school mean values is used in identifying each estimate. Conceptually, this approach removes any source of variation that may be due to the specific school (i.e.,

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

teachers sorting to their preferred schools) but compares outcomes across teachers within schools.

To execute fixed effects here, we aggregate all teachers into a single group and identify those who pass and those who fail with corresponding race-specific indicators. In Table 7, we report the outcomes of this school fixed effects specification (for brevity, we report the estimates using student math achievement only). The magnitude of the estimates are, for the most part, slightly smaller in this table compared with those reported earlier (suggesting that some of the earlier differences may be due to sorting between schools), but the significance and consistency of the findings are the same: The signaling function of the tests operates very differently between teacher demographics for the Praxis II exams.

Nonrandom Attrition

Third, teachers may leave the classroom in nonrandom ways that could bias the classroom data we observe. We test this, similar to Goldhaber (2007), by estimating models restricted to teachers in their first year of teaching. The results are presented in Table 8 (again reporting results for student achievement in math only). The estimates on both the screening and signaling functions appear to be smaller in magnitude (compared with those reported from the full sample of teachers) for Black and male teachers. This suggests that while the test may be somewhat informative for White or female teachers, the information conveyed on Black and male teachers appears to be exceptionally noisy for those brand new to teaching. This conflicts somewhat with the findings presented in Table 3, where we concluded that the screening function of the tests appears to be consistent across demographic groups; with novice teachers only, the difference in the screening function appears to be significant.

Conclusions and Policy Implications

The results presented in this article have several noteworthy dimensions. First, when examining the value of the Praxis exams as a screening device, we find no evidence that they function differently between demographic groups. Our findings show that, regardless of race, these tests do not function as a good screen for teaching effectiveness in reading, but they do function as a reasonable screen for effectiveness in math. This relationship was also analyzed along gender lines, but those results are inconclusive due to low power in the tests. Our findings demonstrate that the screening function of the exams provides fairly uniform information about teachers' ability in the classroom across easily identifiable characteristics; however, this information has limited value since the screen works only when using student achievement in mathematics as the outcome. Moreover, in tests that analyze novice teachers only, we could not detect a significant differential on the

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

Table 7
Licensure Testing Function Estimates With School Fixed Effects in Math

Panel A: Screening Model	Pooled	White/Black (Group 1 = White)	Female/Male (Group 1 = Female)
Passing standard	.028**		
	(.005)		
Group 1 × Pass		.028**	.032**
		(.007)	(.006)
Group 2 × Fail		.000	.021
		(.015)	(.013)
Group 2 × Pass		.025†	.036**
		(.014)	(.008)
Observations	193,462	190,390	193,462
Number of schools	1,114	1,107	1,114
R^2	.67	.68	.67
Panel B: Signaling Model			
Praxis CAE performance	.004*		
	(.002)		
Praxis CIA performance	.009**		
	(.003)		
Group 1 × CAE		.000	.003
		(.002)	(.002)
Group 1 × CIA		.019**	.012**
		(.003)	(.003)
Group 2 × CAE		.028**	.008
		(.005)	(.005)
Group 2 × CIA		−.023**	−.011†
		(.006)	(.006)
Observations	193,462	190,390	193,462
Number of schools	1,114	1,107	1,114
R^2	.68	.68	.68

Note. Robust standard errors in parentheses clustered at the teacher level. Estimates from least squares regressions with the following control variables: indicators on matching relationship, student's race and gender, grade, learning disabilities, free lunch eligibility, limited-English proficiency status, parents' education level, class size, year, and the percentage of minority students in the classroom. Signaling model includes additional teacher-level controls on race and gender, college selectivity, license, National Board certification, educational attainment, experience level, and graduation from a North Carolina–approved education program. CIA = Praxis II Curriculum, Instruction, and Assessment test; CAE = Praxis II Content Area Exercises test.
†$p \leq .10$. *$p \leq .05$. **$p \leq .01$.

screening function in math for Black teachers, further suggesting that these tests may be of limited value. Though we cannot detect differences among

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

Table 8
Novice Teacher Estimates of Licensure Test Functions in Math

Panel A: Screening Model	Pooled	White	Black	Female	Male
Passing current standard	.046*	.060*	.000	.062**	.002
	(.020)	(.024)	(.038)	(.022)	(.044)
Observations	41,449	36,754	4,169	35,734	5,715
R^2	.70	.70	.67	.70	.69
Panel B: Signaling Model					
Praxis CAE performance	.015†	.015†	.009	.011	.046*
	(.008)	(.008)	(.026)	(.008)	(.021)
Praxis CIA performance	.020*	.026*	−.007	.032**	−.031
	(.010)	(.011)	(.024)	(.011)	(.024)
Observations	41,449	36,754	4,169	35,734	5,715
R^2	.70	.70	.67	.70	.69

Note. Robust standard errors in parentheses clustered at the teacher level. Estimates from least squares regressions with the following control variables: indicators on matching relationship, student's race and gender, grade, learning disabilities, free lunch eligibility, limited-English proficiency status, parents' education level, class size, year, and the percentage of minority students in the classroom. Signaling model includes additional teacher-level controls on race and gender, college selectivity, license, National Board certification, educational attainment, experience level, and graduation from a North Carolina–approved education program. CIA = Praxis II Curriculum, Instruction, and Assessment test; CAE = Praxis II Content Area Exercises test.
†$p \leq .10$. *$p \leq .05$. **$p \leq .01$.

demographic groups, we do observe disproportionate numbers of Black teacher candidates among those who fail the licensure standard, which causes some concern. While we find no evidence that this disproportionate failure rate is a result of the test's screening function, it may be influenced by other factors, such as teacher preparation programs and test design. For this reason, we recommend further research into the causes of this performance gap between Black and White teacher candidates.

Second, we find evidence that the tests function differently as signals of quality among Black and male teachers than they do among the dominant groups in the teacher workforce, as measured by student outcomes in both math and reading. Notably, we find that the CIA portion of the Praxis exam is significantly correlated with teacher effectiveness for White and female teachers but not for Black teachers; conversely, Black teacher quality is significantly correlated with the CAE exam, where the same correlations for White or female teachers are considerably smaller. This is important because, though the state itself has no policy of utilizing licensure test performance as quality measures, school or district personnel may use this information at the point of hire as a quantitative measure of quality with

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

low marginal cost. If hiring personnel were to use this information as a proxy quality measure, we would caution against generalizing average test performance to either Black or male teacher candidates, given our findings of evidence showing differences along these lines.

Third, when isolating specific teacher-student interactions, we find evidence that Black teachers have more consistent success than White teachers in teaching minority students, and this matching effect is greatest in magnitude for Black teachers at the lower end of the licensure performance distribution. Moreover, the point estimates suggest that these matching effects are as important as any information conveyed through either the signaling or the screening functions of the tests when it comes to the achievement of minority students.

These findings imply that, were North Carolina to strictly apply its licensure standard and if the current pool of teachers did not change as a result (an admittedly unrealistic assumption), average student outcomes might possibly see a marginal benefit through the exclusion of (on average) lower-scoring teachers (based on the results shown in Table 3). These gains, however, are not likely to be distributed uniformly across students, nor are they certain, since the labor pool will almost certainly change as a result of the more restrictive policy. Furthermore, such a comparison on the basis of the signal value neglects the reality that teachers positively sort across teaching assignments; thus, failing teachers who are excluded from the workforce will likely be substituted with teachers who marginally pass the standard. In fact, the results reported in Table 5 illustrate the potential tradeoff between these marginal average gains and the positive matching effects that benefit minority students. If teachers were paired with same-race students (as commonly results from teachers sorting between schools), these potential average gains would likely come at the cost of adverse effects for minority students. Our tests here are not causal, so we cannot predict this outcome with certainty; however, significant and positive effects are detected among these interacted teacher-student pairings.

To further illustrate this tradeoff, we predict the effect of substituting a Black teacher who fails to meet the North Carolina licensure test cutoff with a White teacher who does meet the cutoff, both teaching in an average Black teacher's classroom. First, we take hypothetical female teachers who score at the 25th percentile of their own-race performance distribution of the combined scores from the Praxis exams. This would mean that the White teacher would have a passing combined score of 325 (standardized to −0.24) while the Black teacher would have a failing score of 304 (standardized to −1.20). We propose a class of 19 students total, comprising 5 White students (26%), 11 Black students (58%), and 3 other-race students (16%).[21] Using the estimates reported previously in Table 5, we predict student outcomes in this hypothetical classroom, holding all else equal. Table 9 reports these predicted outcomes (the predicted values are simply linear predictions

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

Table 9
Comparison Between Predicted Outcomes for Black and White Teachers

Panel A: Differential Classroom Outcomes in Reading	Number in Class	White Teacher	Black Teacher	Difference
White students	5	−.002	.019	−.021
Black students	11	−.001	.054	−.055*
Other race students	3	−.005	.073	−.078*
Total	19	−.002	.048	−.050
Panel B: Differential Classroom Outcomes in Math				
White students	5	−.003	.001	−.005
Black students	11	−.002	.033	−.034*
Other race students	3	−.004	−.011	.008
Total	19	−.003	.017	−.020

Note. Predicted outcomes for hypothetical class, using estimates presented in Table 5. Values represent average student achievement gains for student of given demographic with teacher in 25th percentile of own-race Praxis performance. Estimates in total row are weighted by the number of students in the classroom.
*$p \leq .05$.

given the estimated coefficients in Table 5 and the hypothetical Praxis II scores) and shows that replacing the failing Black teacher with a passing White teacher considerably decreases student outcomes. A class composed of more White students would also decrease in student outcomes in this proposed teacher swap, but the magnitude of the decrease would be considerably smaller. Only in the case of a high-performing White teacher replacing a failing Black teacher (which, on average, seems unlikely) would the net effect of such a switch be nonnegative in this classroom. These average effects, however, mask the fact that minority student achievement decreases the most under this substitution (as shown in the last column). While not causal, this example illustrates the potential tradeoffs inherent in such a policy decision.

We hope that these findings will encourage more substantive discussion on the costs and benefits of teacher licensure tests. As shown above, the tests function in two important ways in North Carolina: as screens and signals of quality. Licensure tests used as screens of quality do appear to provide some limited information about the relative effectiveness of test passers and failers, and the performance of test takers provides additional signal information. In both of the tests' functions, however, the quality of the information conveyed is noisy and varies across different populations of teachers. Thus, we conclude the benefit of testing teachers has some consequential limitations in our data. Compare these with the nontrivial costs associated with licensure

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

testing (as currently practiced in most states), though few of these costs are born directly by states or readily quantifiable.[22] However, the costs to society may be more substantial. For example, prospective teachers incur direct time and money costs in preparing for and taking the exam, and higher labor market entry standards may discourage some prospective teachers from considering teaching (Angrist & Guryan, 2004). Furthermore, as with previous findings, we observe that there are significant numbers of teachers who perform poorly on licensure tests but who are judged to be quite effective (in value-added terms) and that strict adherence to the licensure standard would preclude their participation in the teacher labor market. This is predicted to have notable negative consequences for minority students.[23]

This study sheds new light on the predictive power of licensure test performance among different demographics of teachers and students, but its scope is limited to a focus on observations of individuals teaching in a single state. Thus, as we describe above, we cannot infer much about how licensure testing—and more generally, licensure systems—affects the potential pool of teachers, nor can we infer how these affect cross-state teacher mobility. There is very little empirical research available on these issues, and given the importance of teacher quality in the education process and concerns about teacher supply, these areas are ripe for future research.

Notes

This research is based primarily on confidential data from the North Carolina Education Research Center at Duke University, directed by Clara Muschkin and supported by the Spencer Foundation. The authors wish 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 authors also wish to thank Carol Wallace for editorial assistance, and Mark Long, Hector Cordero, three anonymous referees, and participants at the 2007 APPAM Research Conference for their helpful feedback. The views expressed in this article do not necessarily reflect those of the University of Washington or the study's sponsors. Responsibility for any and all errors rests solely with the authors.

[1]Teacher *effectiveness* and *quality*, as used here and throughout the article, are learning-dependent teaching, as defined by Fenstermacher and Richardson (2005), where student learning forms the output measure of teaching. We clarify this further in the following section.

[2]For more information on teacher-testing policies and states' approaches to setting test cut scores, see Goldhaber (2007).

[3]The programs in California, Alabama, and South Carolina were challenged on the grounds of violation of Title VII of the Civil Rights Act of 1964. Title VII does not specifically outlaw testing procedures that have a significant disparate impact based on the candidate's race, sex, or national origin, but it does require that the uses of the test be valid and consistent with business or educational necessity (Mitchell, Robinson, Plake, & Knowles, 2001).

[4]A ruling in 1985 (*Margaret Allen, et al. v. Alabama State Board of Education, et al.*) resulted in a moratorium on teacher testing that lasted from the early 1990s into this decade (in 2002, Alabama reinstated teacher testing with the Alabama Basic Skills Test and the Alabama Prospective Teacher Test for initial certification). In a 1989 case (*Richardson v. Lamar County Board of Education*), the plaintiff

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

claimed that the school district wrongfully refused to renew her teaching contract because the certification exams required for the position had an adverse racial impact. The court ruled that she could recover on her claim of "disparate impact." Unlike in the other states, the tests were found to be content invalid, meaning that a passing test on the score did not demonstrate the necessary knowledge to be a minimally competent teacher.

[5]An older literature that relies on more aggregated data also tests to find a positive relationship between teacher testing and student achievement (Ferguson, 1991; Ferguson & Ladd, 1996; Strauss & Sawyer, 1986).

[6]We do not address these issues here; however, teacher testing may have other important effects on the teacher workforce besides screening out low-scoring individuals. For example, testing increases the cost of labor market participation (Friedman & Kuznets, 1945), may increase wages to teachers (Angrist & Guryan, 2008), and could discourage otherwise qualified people from becoming teachers (Angrist & Guryan, 2004; Hanushek & Pace, 1995), and teacher test performance may provide a signal of employee quality that could influence public school hiring decisions.

[7]This may arise for a number of reasons, such as tests having a cultural bias in either direction, where test takers are evaluated on knowledge that is more prevalent in one group than the other and not necessarily related to job performance (Jencks, 1998).

[8]It is important to note that meeting the licensure testing standard is generally one of several prerequisites to attaining a license to teach.

[9]The quote is

> The lack of an exact value for the highest score obtainable follows from the fact that the Praxis™ 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 Educational Testing Service website posted replies to questions: http://www.ets.org/portal/site/ets/menuitem.2e37a093417f63e3aa77b13bc3921509/?vgnextoid=a2912d3631df4010VgnVCM10000022f95190RCRD &vgnextchannel=57ec253b164f4010VgnVCM10000022f95190RCRD.

[10]The North Carolina Education Research Data Center, which assembles and distributes the data we use in this study, notes that teachers monitoring their own classes' exams is standard practice; in the data codebook, the center writes, "An instructor is associated with the group of students who took the exam together. *In many cases, this instructor will be that group's teacher, but this is not necessarily the case*" (emphasis added). Restricting the sample to self-contained classrooms prevents us from attributing all students' learning gains to a single teacher when multiple teachers may have taught a single classroom. Imposing this restriction, however, removes most sixth graders, as the great majority of sixth-grade students in North Carolina are enrolled in middle schools (they make up, in total, less than 2% of the student observations in our analysis sample). For a discussion of the changes to a sample resulting from relaxing the restriction to self-contained classrooms, see Clotfelter, Ladd, and Vigdor (2003).

[11]The North Carolina teacher-student data have been utilized for a good deal of research that has made its way into well-respected peer-reviewed economics and education journals (e.g., Clotfelter et al., 2005, 2006; Goldhaber, 2007; Goldhaber & Anthony, 2006; J. Rothstein, forthcoming).

[12]Students are not tested before the third grade, and they almost always switch teachers for grades above the sixth; hence, we exclude those that do not meet the data restrictions as described. The result of this is that the large majority of student observations that our data sample are from Grades 4 and 5.

[13]The results we report here are robust to changes in these class size restrictions. We also performed the analysis (not reported here) on a sample where class size was

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

restricted to having 8 to 27 students (corresponding to removing the upper and lower 2.5 percentiles) and a sample where no class size restrictions were imposed. Our findings here were not substantively affected.

[14]The names of the Praxis tests are slightly confusing. The Curriculum, Instruction, and Assessment test is actually a multiple-choice exam covering content knowledge of elementary skills. The Content Area Exercises test, on the other hand, is a free-response essay exam, prompting test takers with different classroom teaching situations (e.g., detecting and diagnosing problems in student work) and requiring the teacher to lay out his or her strategy in response to the situation.

[15]In these tables and in the analyses that follow, when comparing teachers across racial characteristics, we consider only Black and White teachers, as the sample of Hispanic teachers or teachers from other ethnic backgrounds is too small for reliable inference. When comparing across gender characteristics, we include all teachers of that gender in the analysis without regard to racial background. We do not interact race and gender for reasons of low power.

[16]Again, the Praxis test scores are normalized against the population of teachers taking the same test in the same year, not only against the teachers in the sample. We omitted from the sample those who were teaching grades K-3 (there are no test data available for students in those grades) and those teachers who were not linked to a classroom of students.

[17]We wish to clarify that a teacher's pass-fail status can change over time, depending on his or her retakes of the exam; hence, we report unique teachers who have ever taught while failing the standard. Also, some teachers may have failed the exam and passed on a successive retake prior to actually teaching in a classroom; these teachers are not included in this count, as they never taught while failing the standard. If we consider the lowest test scores of all teachers, regardless of whether they taught while failing, we find the failure rates of unique teachers in our sample on the current standard are 0.56 for Black teachers, 0.14 for White teachers, 0.27 for male teachers, and 0.18 for female teachers.

[18]To have "full licensure" in North Carolina a person must have a Standard Professional 1 (SP1) or a Standard Professional 2 (SP2) license. The SP1 is granted to individuals who have completed a state-approved teacher education program or have completed another state's approved route to licensure and met the federal requirements to be designated "Highly Qualified" and earned a bachelor's degree from a regionally accredited college. The SP2 license is granted to in-state teachers with 3 or more years of teaching experience. Also, teachers from another state who have been fully licensed in that state, are "Highly Qualified," have 3 or more years of teaching experience, and who either meet North Carolina's Praxis testing requirements or have National Board Certification are issued the SP2 license.

[19]The regression equation is $A_{i,j,t}=\alpha A_{i,t-1}+\beta_1 \text{STUDENT}_{i,j,t}+\delta_j+\epsilon_{i,j,t}$, where δ_j is the fixed teacher effect on student outcomes for all students under teacher j in the sample. This teacher effect is estimated for each teacher and then averaged across all teachers of a similar demographic. Teacher characteristics such as race, gender, years of experience, and so on, are not controlled for separately but are incorporated into the estimate on teacher effectiveness.

[20]In addition to estimating the models presented here, we estimated intermediary models between the pooled model (which restricts the estimated intercepts to be equal across all teachers) and those models where demographic differences are mutually exclusive. In particular, we estimated models that shared estimates across all variables except for the gains to passing the licensure standard for each subgroup. While this did prove to fit the data better than the pooled model, a series of tests assessing the goodness of fit to the various models gave us reason to believe that the demographic differences among teachers were in fact structural differences and not just different constant terms to each race or gender; hence, we present the results of the models estimated exclusively.

[21]As documented earlier (note 16), 56% of Black teachers in our sample have minimum Praxis scores below the cutoff, so the hypothetical Black teacher that fails the standard is not an outlier. The proposed class matches the average percentage of minority students (.74) in classes taught by Black teachers in our sample (see Table 2).

[22]The direct, easily quantifiable costs that states incur are likely to be relatively small in the case of North Carolina. For example, the Praxis tests we analyze are developed on

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

a national level by a third party (thus, low cost to the state), and the administration and enforcement of the licensure standard are also low cost (performance on the exam is a single quantitative measure in North Carolina, which is easy to track and judge).

[23]We wish to note a limit of our results: The outcomes that we use throughout this article, student achievement in either reading or math, are also outcomes of standardized tests. These tests, like the Praxis tests that we investigate here, may signal student learning differently across student demographic groups, which may limit our usefulness in generalizing our results to the entire student population. Without an additional outcome measure to enable us to calibrate the predictive validity of student test scores on other outcomes across demographic groups, we cannot say for sure how well these significant differences in student test scores correlate with actual differences in student outcomes that are of policy interest (such as successfully graduating high school or enrolling in college). Further investigation of this related issue is beyond the scope of the current study.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135.

Angrist, J. D., & Guryan, J. (2004). Teacher testing, teacher education, and teacher characteristics. *American Economic Review*, *94*(2), 241–246.

Angrist, J. D., & Guryan, J. (2008). Does teacher testing raise teacher quality? Evidence from state certification requirements. *Economics of Education Review*, *27*(5), 483–503.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2003). Segregation and resegregation in North Carolina's public school classrooms: Do Southern schools face rapid resegregation?*North Carolina Law Review*, *81*(4), 1463–1512.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review*, *24*, 377–392.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, *41*(4), 778–820.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). *How and why do teacher credentials matter for student achievement?* (NBER Working Paper No. 12828). Cambridge, MA: National Bureau of Economic Research.

Coleman, J. S. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Department of Health, Education, and Welfare.

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *The Review of Economics and Statistics*, *86*(1), 195–210.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender Matter? *The American Economic Review*, *95*(2), 158–165.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources*, *42*(3), 528–554.

Dilworth, M. E. (1986). Teacher testing: Adjustments for schools, colleges, and departments of education. *The Journal of Negro Education*, *55*(3), 368–378.

Dometrius, N. C., & Sigelman, L. (1988). The cost of quality: Teacher testing and racial-ethnic representativeness in public education. *Social Science Quarterly*, *69*(1), 70–82.

Ehrenberg, R. G., Goldhaber, D. D., & Brewer, D. J. (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.

Farrell, E. J. (1990). On the growing shortage of Black and Hispanic teachers. *The English Journal*, *79*(1), 39–46.

Fenstermacher, G. D., & Richardson, V. (2005). On making determinations of quality in teaching. *Teachers College Record*, *107*(1), 186–213.

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

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). Can schools narrow the Black-White test score gap. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 318–374). Washington, DC: The Brookings Institution.

Ferguson, R. F., & Ladd, H. F. (1996). How and why money matters: An analysis of Alabama schools. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 265–298). Washington, DC: The Brookings Institution.

Fordham, S., & Ogbu, J. U. (1986). Black students' school success: Coping with the "burden of 'acting White.'" *The Urban Review*, *18*(3), 176–206.

Friedman, M., & Kuznets, S. S. (1945). *Income from independent professional practice*. New York: National Bureau of Economic Research.

Gitomer, D. H. (2007). *Teacher quality in a changing policy landscape: Improvements in the teacher pool*. Princeton, NJ: Educational Testing Service.

Gitomer, D. H., & Latham, A. S. (2000). Generalizations in teacher education: Seductive and misleading. *Journal of Teacher Education*, *51*(3), 215–220.

Goldhaber, D. (2007). Everyone's doing it, but what does teacher testing tell us about teacher effectiveness? *Journal of Human Resources*, *42*(4), 765–794.

Goldhaber, D., & Anthony, E. (2006). Can teacher quality be effectively assessed? National Board certification as a signal of effective teaching. *The Review of Economics and Statistics*, *89*(1), 134–150.

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy*, *100*(1), 84–117.

Hanushek, E. A., & Pace, R. R. (1995). Who chooses to teach (and why)? *Economics of Education Review*, *14*(2), 101–117.

Holmlund, H., & Sund, K. (2008). Is the gender gap in school performance affected by the sex of the teacher? *Labour Economics*, *15*(1), 37–53.

Irvine, J. J. (1988). An analysis of the problem of disappearing Black educators. *The Elementary School Journal*, *88*(5), 503–513.

Irvine, J. J. (1992). Making teacher education culturally responsive. In M. E. Dilworth (Ed.), *Diversity in teacher education: New expectations* (pp. 79–92). San Francisco: Jossey-Bass.

Jencks, C. (1998). Racial bias in testing. In C. Jencks & B. R. Phillips (Eds.), *The Black-White test score gap* (pp. 55–85). Washington, DC: Brookings Institution Press.

King, S. H. (1993). The limited presence of African-American teachers. *Review of Educational Research*, *63*(2), 115–149.

Kirby, S. N., Naftel, S., & Berends, M. (1999). *Staffing at-risk school districts in Texas: Problems and prospects*. Santa Monica, CA: Rand.

Kleiner, M. M., & Kudrle, R. T. (2000). Does regulation affect economic outcomes? The case of dentistry. *Journal of Law & Economics*, *43*(2), 547–582.

Lavy, V. (2004). *Do gender stereotypes reduce girls' human capital outcomes? Evidence from a natural experiment* (NBER Working Paper No. 10678). Cambridge, MA: National Bureau of Economic Research.

Margaret Allen, et al. v. Alabama State Board of Education, et al., 81-T-697-N (United States District Court, M. D. Alabama, Northern Division 1985).

Medley, D. M., & Quirk, T. J. (1974). The application of a factorial design to the study of cultural bias in general culture items on the National Teacher Examination. *Journal of Educational Measurement*, *11*(4), 235–245.

Mitchell, K. J., Robinson, D. Z., Plake, B. S., & Knowles, K. T. (Eds.). (2001). *Testing teacher candidates: The role of licensure tests in improving teacher quality*. Washington, DC: National Academy Press.

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016

North Carolina Department of Public Instruction. (2007a). *Licensure*. Retrieved February 19, 2009, from http://www.ncpublicschools.org/licensure/

North Carolina Department of Public Instruction. (2007b). *North Carolina standard course of study*. Retrieved February 19, 2009, from http://www.ncpublicschools.org/curriculum/ncscos

Ogbu, J. U. (1974). *The next generation: An ethnography of education in an urban neighborhood*. New York: Academic Press.

Ogbu, J. U. (1978). *Minority education and caste: The American system in cross-cultural perspective*. New York: Academic Press.

Ogbu, J. U. (1992). Understanding cultural diversity and learning. *Educational Researcher*, *21*(8), 5–14.

Richardson v. Lamar County Board of Education, 729 F. Supp. 806 (United States District Court, M.D. Alabama, Northern Division 1989).

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (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., & Mead, S. (2004). Back to the future: The history and politics of state teacher licensure and certification. In F. M. Hess, A. J. Rotherham, & K. Walsh (Eds.), *A qualified teacher in every classroom: Appraising old answers and new ideas* (pp. 11–47). Cambridge, MA: Harvard Education Press.

Rothstein, D. S. (1995). Do female faculty influence female students' educational and labor market attainments? *Industrial & Labor Relations Review*, *48*(3), 515–530.

Rothstein, J. (forthcoming). Student sorting and bias in value added estimation: Selection on observables and unobservables. *Education Finance and Policy*.

Sanders, W. L., Ashton, J. J., & Wright, S. P. (2005). *Comparison on the effects of NBPTS certified teachers with other teachers on the rate of student academic progress*. Cary, NC: SAS Institute.

Stigler, G. J. (1971). The theory of economic regulation. *Bell Journal of Economics and Management Science*, *2*(1), 3–21.

Strauss, R. P., & Sawyer, E. A. (1986). Some new evidence on teacher and student competencies. *Economics of Education Review*, *5*(1), 41–48.

U.S. Department of Education. (2006). *The secretary's fifth annual report on teacher quality: A highly qualified teacher in every classroom*. Retrieved February 19, 2009, from https://title2.ed.gov/Title_II_06.pdf

Vars, F. E., & Bowen, W. G. (1998). Scholastic Aptitude Test scores, race, and academic performance in selective colleges and universities. In C. Jencks & B. R. Phillips (Eds.), *The Black-White test score gap* (pp. 457–479). Washington, DC: Brookings Institution Press.

Villegas, A. M., & Clewell, B. C. (1998). Increasing the number of teachers of color for urban schools: Lessons from the pathways national evaluation. *Education and Urban Society*, *31*(1), 42–61.

Zapata, J. T. (1988). Early identification and recruitment of Hispanic teacher candidates. *Journal of Teacher Education*, *39*(1), 19–23.

Zimpher, N. L., & Ashburn, E. A. (1992). Countering parochialism in teacher candidates. In M. E. Dilworth (Ed.), *Diversity in teacher education: New expectations* (pp. 40–62). San Francisco: Jossey-Bass.

Manuscript received August 15, 2008
Revision received July 17, 2009
Accepted August 11, 2009

Downloaded from http://aerj.aera.net at PENNSYLVANIA STATE UNIV on May 8, 2016