

National Board Certification and Teacher Effectiveness: Evidence from a Random Assignment Experiment

By

Steven Cantrell
Learning Point Associates

Jon Fullerton
Harvard Graduate School of Education

Thomas J. Kane
Harvard Graduate School of Education

Douglas O. Staiger
Dartmouth College

November 14, 2007

This analysis was supported by the Spencer Foundation. Initial data collection was supported by a grant from the National Board on Professional Teaching Standards to the Urban Education Partnership in Los Angeles. At the outset of this project, Cantrell was Chief Research Scientist in the Program Evaluation and Research Branch of the Los Angeles Unified School District (LAUSD), Fullerton was at the Urban Education Partnership (UEP) and Kane was at UCLA. The authors wish to thank a number of current and former employees of LAUSD, including Ted Bartell, Jeff White, Glenn Daley, Jonathan Stern and Jessica Norman. From the Urban Education Partnership, Susan Way Smith helped initiate the project and Erin McGoldrick oversaw the first year of implementation. An external advisory board composed of Eric Hanushek, Daniel Goldhaber and Dale Ballou provided guidance on initial study design. Jeffrey Geppert helped with the early data assembly.

Abstract:

The National Board for Professional Teaching Standards assesses teaching practice based on videos and essays submitted by teachers. We compared the performance of classrooms of elementary students in Los Angeles randomly assigned to NBPTS applicants and to comparison teachers. The students assigned to highly-rated applicants outperformed those in the comparison classrooms by more than those assigned to poorly-rated teachers. Moreover, we compare experimental and non-experimental estimates of the predictive value of NBPTS ratings (for the *same* sample of teachers in earlier years as well as for a non-experimental sample). The estimates were similar to those based on the experiment. We make a number of suggestions for improving the predictive power of the NBPTS scaling process.

I. Introduction and Motivation

Research in a variety of school districts and states has suggested that there are large and persistent differences in teachers' impacts on students' academic achievement. However, there is much less agreement on the traits and teaching practices that underlie those differences. Over the past decade, many districts and states have begun to rely on the National Board for Professional Teaching Standards (NBPTS) to identify their most effective teachers by scoring videotapes and essays. In this paper, we evaluate the ability of the NBPTS to identify those teachers with the biggest impact on student achievement.

Broadly speaking, there are two approaches to assessing teacher performance: estimating impacts on student achievement directly (using longitudinal test score data on teachers and students to generate so-called "value-added" estimates) and observing and rating teachers' classroom practice (without reference to student achievement). The NBPTS process is an example of the latter. To apply for certification from the NBPTS, a teacher must submit a portfolio of their work (including examples of written feedback to students, a self-assessment of effectiveness and videotaped examples of lessons) and respond to six essay questions at an assessment center. The NBPTS has developed a method for scoring those submissions against a set of standards they developed.

In this paper, we test whether the scores issued by the NBPTS are related to teacher impacts on student achievement. We also explore the gains to be made from combining the two approaches-- using *both* prior value-added estimates and practice-based approaches to identify effective teachers.

Several recent papers have assessed the validity of NBPTS certification in identifying those teachers with the largest estimated impacts on student achievement. (Goldhaber

and Anthony (2004), Cavaluzzo (2006), Vandervoort et. al. (2004), Clotfelter, Ladd and Vigdor (2006), Sanders, Ashton and Wright (2005) and Harris and Sass (2006)). Such research has generally found differences in student achievement impacts of .05 to .10 standard deviations between certified teachers and unsuccessful applicants.

Although several earlier papers have studied changes in teacher's impacts before, during and after the NBPTS application process, we are primarily interested in the ability of the NBPTS to recognize effective teachers, not the impact of the process on the effectiveness of applicants. As a result, we compare the performance of those *ever identified* as being a NBPTS certified teacher to those ever rated poorly in the NBPTS process. We extend the earlier research in a number of important ways:

First, unlike earlier studies, we use random assignment to compare the student achievement impacts of NBPTS applicants (both certified and uncertified) to non-applicants working in the same schools and grade levels. For this study, the NBPTS identified all of those who had applied for certification from within the zip codes in the Los Angeles region. For 99 of such NBPTS applicants, LAUSD identified comparison teachers teaching in the same school, grade and calendar track to serve as comparisons. The district then asked their principals to identify two classrooms that they would be willing to assign to either teacher, and randomly assigned the classrooms to each one of the teachers in each pair. We compare their performance at the end of the year.

Second, we use information on each applicant's NBPTS scaled score (not just whether the candidates achieved certification) to test whether the score is related to teacher impacts. All prior studies have used simple dichotomous comparisons—either comparing those certified by NBPTS to unsuccessful applicants or to all others (a

combination of non-applicants and unsuccessful applicants). However, such comparisons conflate the information contained in the scaled score with the *distribution* of scores of applicants above and below the cut-off. Because the cut-off for NBPTS certification is drawn near the mean of the scaled score distribution (roughly half of those who take the exam fail), there are large numbers of applicants with scores right above and right below the cut-off. (In fact, the difference in mean scaled scores between successful and unsuccessful applicants is minimized at the current cut-off.) We test the predictive value of the continuous scaled score, not just whether or not an applicant achieved certification.

Third, we test the predictive value of each of the 10 sub-scores that make up the NBPTS' scaled score. Lacking any student achievement data with which to validate their sub-scores, the NBPTS chose these weights based on their own professional judgment, without reference to student achievement impacts. We revisit those judgments by validating against student achievement impacts (essentially including each of the sub-scores separately and testing the NBPTS weighting).

Fourth, we compare experimental and non-experimental estimates of the impact of NBPTS teachers in LAUSD. While the random assignment occurred during the 2003-04 and 2004-05 school years, we also have longitudinal data for the same set of teachers during the 1999-2000 through 2002-2003 school years-- when the same teachers were assigned to classrooms in the usual manner. Moreover, many NBPTS applicants were not chosen for randomization. We compare the estimates for the experimental *sample* during the experimental *period* (spring 2004 and 2005) to three different non-experimental estimators: for the *non-experimental sample* during the *experimental*

period (2004-2005); for the *non-experimental sample* during the *pre-experimental period* (2000-2003); and, for the *experimental sample* during the *pre-experimental period*.

Finally, we test the predictive power of the NBPTS scaled score while controlling for a non-experimental “value-added” estimate from prior years.

We report four primary findings: First, in the experiment, we find that those who achieved certification were not statistically significantly more effective than non-applicants; but un-successful applicants were less effective than non-applicants. The difference in impacts between successful and unsuccessful applicants was statistically significant—with non-applicants somewhere in between the successful and unsuccessful applicants. Second, our non-experimental estimates are similar, although somewhat smaller in magnitude, than the experimental estimates. Third, the NBPTS’s ability to predict student achievement impacts could be roughly doubled, simply by re-weighting the 10 components in calculating the scaled score. Finally, for individual teachers, the non-experimental estimates of their value-added in the years prior *to* random assignment had considerable predictive power in predicting student achievement during the experiment.

The remainder of the report proceeds as follows. We provide some background on the NBPTS application process and their scoring. Next, we review the recent literature on the relationship between NBPTS certification and student achievement and describe the process by which the experimental sample was chosen. Then, we describe our estimation strategy present the results from both the experimental and non-experimental samples.

II. The NBPTS Application Process

The process of becoming a National Board Certified Teacher is time-intensive and can take from three months to several years. Candidates are required to submit a portfolio and to complete a series of written exercises at a testing site. The portfolio entries include written commentaries on student work, video tapes of and commentaries on classroom lessons, and evidence of engagement with the school community. The Assessment Center exercises are short (30 minute) essay questions designed to test the candidate's pedagogical content knowledge.

The four portfolio entries and six assessment center essays are each scored on a four-point scale.¹ The raw score for each of the 10 items is weighted to generate a scaled score (the sum of the 10 weighted sub-scores), which has ranged between 87 and 437.² The candidate is required to achieve a scaled score of 275 in order to receive certification.³

NBPTS currently provides certificates in 24 different areas, varying by developmental level (e.g., early childhood, early adolescence) and content area (e.g., art, mathematics, generalist). Given our focus on elementary schools, the vast majority of applicants were drawn from two areas: early childhood generalists (who work with students aged 3 through 8) and middle childhood generalists (who work with students aged 7 through 12).

¹ Candidates who began their application process prior to 2002 completed six portfolio entries and four Assessment Center exercises. The score can include pluses and minuses – so actual entry results range from .75 (1-) to 4.25 (4+).

² The weights add up to 100. Finally, a constant of 12 is added to the score to generate a final scaled score between 87 and 437.

³ Candidates must also complete all ten entries to receive certification even though it is theoretically possible to have a scaled score higher than 275 without completing all of the entries. (NBPTS 2006a)

In California, candidates are required to hold a teaching credential for three years before they apply for National Board certification.⁴ Currently, the cost to apply is \$2,500, but was \$2,300 during the period of the study. Candidates could apply for a subsidy from the state of California to pay half of this fee; additional subsidies are sometimes available from the National Board as well as other organizations. Several different organizations provided classes in Los Angeles to help applicants complete the process.

Table 1 presents an overview of components of the assessment for the middle childhood generalist certificate. Importantly, the process changed for those applying for the first time in 2002. (Those who had started the process prior to 2002 were scored under the old system.) In 2002, several portfolio entries were combined and the number of Assessment Center exercises was raised from four to six. Table 2 shows the changes from the “old” to the “new” certification processes.

Finally, candidates who do not attain certification in the first year they apply are allowed to “bank” their scores for up to 24 months. During this period, candidates may retake any single one of the portfolio entries or assessment center exercises on which they received a score of less than 2.75.⁵ The retake score replaces the original score, whether or not it is higher than the original score. Total scaled scores are then recalculated and National Board Certification is awarded to those whose new scores allow them to achieve higher than 275 scaled score points. Although roughly half of applicants fail in their first try, approximately two-thirds of initial applicants eventually pass when retakes are considered.

⁴ Holding an intern credential or emergency teaching permit does not count towards this requirement (NPBTS 2006a, 3-4).

⁵ Candidates are charged \$350 for each entry or exercise they retake (NBPTS 20006b, 33,34).

National Board Teachers in the Los Angeles Unified School District

The Los Angeles Unified School District (LAUSD) is the second largest school district in the nation. In 2005, LAUSD enrolled over 727,000 K-12 students and employed over 37,000 regular teachers. The state of California and LAUSD created a number of incentives to encourage the growth National Board Certified Teachers (NBCTs) in Los Angeles. Until 2003-04, the state provided a one time \$10,000 award to teachers who successfully completed the certification process. Although this one-time award was eliminated, the state continues to provide \$5,000 per year for four years to NBCT teachers who teach in “high-priority” schools, based on their performance on the state tests.⁶ Four-fifths (80%) of LAUSD students attend such “high-priority” schools.

As part of its collective bargaining agreement, LAUSD supplements the state incentives. NBCTs receive an ongoing 7.5% increase on their base salary for their accomplishment. In addition, if a NBCT provides the District with 92 hours of “service” (generally professional development or mentoring activities), he or she will receive an additional 7.5% pay increment.

In total, the financial incentive to gain National Board Certification can be quite substantial. In light of these financial incentives, LAUSD witnessed a robust response in the the number of teachers applying to the NBPTS for certification. As of 2004, 1790 LAUSD teachers had applied for NBPTS certification with 1129 having achieved this certification. In fact, LAUSD has more National Board Certified Teachers than any other district—in terms of absolute numbers of candidates. By 2005, the district was spending roughly \$7 million on the program annually.

⁶ The state defines a high-priority school as a school in the bottom half of the State Academic Performance Index Rankings.

III. **Literature Review**

Previous research studying the link between National Board Certification and student academic outcomes has had two major limitations. First, all of the previous studies have relied upon observational data. As a result, much of the discussion about the impact of NBCTs has been bogged down in debates about the appropriate specification of the empirical models. Second, all of the previous studies have looked solely at whether NBCTs are more effective as a *group* than other teachers. None of the prior studies have examined whether the scaled score and individual exercise scores are effective in predicting teacher impacts on student achievement.

Research on the effectiveness of NBCTs

The findings of the studies evaluating the impact of National Board certified teachers have been mixed. (The earlier studies are briefly summarized in Table 3.) Two early studies in this literature (Goldhaber and Anthony 2005, Cavalluzzo 2004) found that NBCT's were somewhat more effective at raising student achievement than other teachers who did not apply for certification. They also found that NBCT's were even more effective than unsuccessful applicants. The impact of NBCT's was significant but relatively modest in both studies (.05 in math for Goldhaber and Anthony, .07 in math for Cavaluzzo).

A later study by Sanders et al. (2005) called these findings into question, noting that neither of the previous analyses properly accounted for teacher-level random effects—(that is, classroom-level or teacher-level variation in impacts on student

achievement). Even if there is a difference in their mean effectiveness, we might not expect all National Board certified teachers to outperform all non-applicants (i.e. there might be a teacher-level random effect generating a distribution of outcomes in both groups). Using their preferred models, Sanders et al. found similar effect sizes to those reported by Goldhaber and Anthony-- .05 to .07 in math. However, the size of the standard errors dramatically increased after allowing for teacher random effects, with the result that most of the estimates in the Sanders et al. study were found to be statistically insignificant.

Harris and Sass (2007) included both student and school fixed effects to their analysis of NBCT's in the state of Florida. They found that National Board certification does indicate higher teacher productivity in some grades, subjects and years, but not in all. In addition, they found different results depending on whether they use the Florida Comprehensive Achievement Test or the SAT-9 as the response variable.

Finally, Clotfelter, Ladd, and Vigdor in their own analysis of data from North Carolina do find a statistically significant impact on a student's achievement of having been assigned a National Board certified teacher. Their comparison group was all other teachers—whether or not they applied for National Board certification, and as a result their estimate is somewhat smaller .02-.03 standard deviations in math.

Research on Other Practice-Based Assessments

There have been a number of other studies on the relationship between objective and subjective measures of teacher performance (Daley 2006, Gallagher 2004, Jacob & Lefgren 2005, Kimball et al. 2004, Milanowski 2004). This work has attempted to

discern the relationship between evaluators' ratings of teachers and those same teachers' actual impact upon student achievement.

The National Board process is somewhat unique in that it combines high stakes (i.e., significant pay differentials) with an evaluation process that is carried out by a neutral third party (i.e., the National Board as opposed to principals or other supervisors). Labor representatives typically worry that high stakes performance evaluations given by supervisors will be vulnerable to arbitrary favoritism and discrimination on the part of the evaluators. Indeed, while Jacob and Lefgren (2005) find that principals can identify teachers with the largest and smallest impacts on student achievement, they also find that principals discriminate in favor of teachers they have a good relationship, as well as by gender and tenure status.

On the other hand, the arms-length evaluation given by the National Board has some disadvantages. First, the information submitted in the portfolio entries is largely self-reported by the candidate. The National Board cannot know how much coaching went on before the video was selected, how many times lessons have been taught (or re-taught), or the number of "failed" lessons videotaped prior to the submitted tape. Second, the National Board evaluators have no real access to any "local knowledge" of the school. Such questions as "Has the applicant been assigned particularly high performing or low performing students?" are unanswerable. Third, evaluators do not have direct access to parent, colleague, or principal opinion regarding the performance of the teacher in the school. In addition, the NBPTS process is costly in terms of teachers' and evaluators' time.

IV. Experimental Assignment

The experimental portion of the study took place over two school years: 2003-04 and 2004-05. The NBPTS provided the research team with a list of all past and present National Board applicants that lived in the Los Angeles area (identified by zip code) at the time of application. LAUSD matched this list with their current employees, allowing the team to identify those teachers still employed by the District.

The sample population was restricted to grades two through five, since students in these grades typically are assigned a single instructor for all subjects. Once the National Board applicants were identified, the study team identified a list of comparison teachers in each school. Comparison teachers had to teach the same grade and be part of the same calendar track as the National Board Applicants. In addition, the NBPTS requires that teachers have at least three years of experience before application. Since prior research has suggested that teacher impacts on student achievement grow rapidly during the first three years of teaching, we restricted the comparison sample to those with at least three years of teaching experience.

School principals were sent a letter from the District's Chief of Staff that requested their participation in the study and gave details on the process. These letters were subsequently followed up with phone calls from the District's Program Evaluation and Research Branch (PERB). However, school participation in the study was voluntary. If a principal agreed to participate, then PERB staff confirmed the appropriate comparison teacher with the principal. The comparison teacher selected by the study team could be inappropriate for many reasons. First, the data that the research team used to generate appropriate comparison teachers was based on the prior year's data. If either

the applicant or comparison teacher changed grade, track, or employment status between years, the research team comparison selection would become invalid. Second, many elementary classes in LAUSD, as in most other districts, are not interchangeable. For instance, basic English Learners may be concentrated in a class with a teacher that has experience and training in working with English Learners. As it would be inappropriate to “switch” such teachers, the comparison teachers were dropped and another one found, if possible.

Once the comparison teacher was chosen, the principal was asked to choose a date upon which the random assignment of rosters to teachers would be made. (Principals either sent PERB rosters or already had them entered into LAUSD’s student information system). Typically, principals wanted this to be as late as possible in the summer but before teachers arrived back at school. This timing would minimize the amount of enrollment change while not interfering with teachers planning. On the chosen date, LAUSD’s PERB in conjunction with the LAUSD’s School Information Branch randomly chose which rosters to switch and executed the switches at the Student Information System at the central office. Principals were then informed whether or not the roster switch had occurred. Ninety-nine valid pairs of teachers were generated for the experimental portion of the study this way.

Once the roster switches had occurred, no further contact was made with the school. Some students presumably later switched between classes. However, 85 percent of students remained with the assigned teacher at the end of the year. Teacher and student identifiers were masked by the district to preserve anonymity.

The National Board provided the research team with additional information including scaled scores, scores on individual entries and exercises, and application dates for all NBCTs in the LA area. LAUSD then linked all of these scores to the masked identifiers to allow the research team to complete its analysis.

V. Data

During the 2002-03 academic year, the Los Angeles Unified School District (LAUSD) enrolled 746,831 students (kindergarten through grade 12) and employed 36,721 teachers in 689 schools scattered throughout Los Angeles County. There were 429 elementary schools in the district.⁷

We use test score data from the spring of 1999 through the spring of 2005. Between the spring of 1999 and the spring of 2002, the Los Angeles Unified School District administered the Stanford 9 achievement test. Under state regulations, exemptions were not granted to students with disabilities or poor English skills. In May 2002, test scores were available for 90 percent of students enrolled in grades 2 through 5. In the Spring of 2003, the district (and the state) switched from the Stanford 9 to the California Achievement Test. During the 2003-2004 and 2004-2005 academic years (the experimental period), the district used a third test—the California Standards Test. For each test and each subject, we standardized by grade and year.

Although there was considerable mobility of students within the school district (9 percent of students in grades 2 through 5 attended a different school than they did the previous year), the geographic size of LAUSD ensured that most students remained

⁷ Student enrollment in LAUSD exceeds that of 29 states and the District of Columbia.

within the district even if they moved. Conditional on having a baseline test score, we observed a follow-up test score for 90 percent of students in the following spring.

We observed snapshots of classroom assignments in the fall and spring semesters. In both the experimental and non-experimental samples, our analysis focuses on “intention to treat” (ITT), using the characteristics of the teacher to whom a student was assigned in the fall. As we mention below, classroom switching was not very common in the experimental sample, so that instrumental variables estimates of the treatment effect (using assigned teacher as an instrument for actual teacher) are never more than 20% larger than those we report.

We also obtained administrative data on a range of other demographic characteristics and program participation. These included race/ethnicity (hispanic, white, black, other or missing), indicators for those ever retained in grade, designated as Title I students, those eligible for Free or Reduced Price lunch, those designated as homeless, migrant, gifted and talented or participating in special education. We also used information on tested English language Development level (level 1-5). In many specifications, we included fixed effects for the school, year, calendar track and grade for each student.⁸

We dropped those students in classes where more than 20 percent of the students were identified as special education students. Finally, we dropped classrooms with extraordinarily large (more than 36) or extraordinarily small (less than 10) enrolled students (3 percent of students with valid scores).

⁸ Because of overcrowding, LAUSD operates a number of schools on a year-round calendar—with students on up to four different schedules rotating their attendance throughout the year, which we refer to a calendar track.

We provided a list of zip codes in the Los Angeles area. The NBPTS then provided a list of all applicants who had LA zip codes at the time of the application. For each applicant, we obtained their National Board status (passed, failed or withdrew) along with their overall scaled score and score on each of the ten sub-scores. For individuals who retook some sections, we obtained both their initial and final scores.

We also obtained snapshots of all district employees from 1994 through 2005. Therefore, for teachers who were hired since 1993, we observed actual years of teaching experience since the time of hiring. Our sample of teachers who did not apply to the National Board is limited to teachers with at least 3 years of experience, to avoid comparison of National Board applicants to novice teachers (who are known to be less effective at improving student test scores).

VI. Empirical Methods

The experimental sample included 99 pairs of teachers teaching in the same school, grade and year. Each pair had one teacher who was a National Board applicant and one teacher who was a non-applicant with at least 3 years of teaching experience. Within each pair, class rosters were randomly assigned. The non-experimental sample included all remaining National Board applicants who were teaching in grades 2-5, along with all other teachers with at least 3 years of experience teaching in the same school-grade-year as a National Board teacher. In the non-experimental sample, class rosters were assigned by the principal in the usual manner.

Estimating Impacts of National Board Applicants on Student Achievement

We tested whether National Board certification was related to teacher impacts on student achievement using two basic specifications. The first specification was as follows:

$$(1) \quad s_{i,yr} = \lambda_1 Cert_j + \beta_g S_{i,yr-1} + \phi X_{i,yr} + \gamma \bar{X}_{j,yr}^c + \delta_{s,g,tr,yr} + \varepsilon_{i,yr}$$

The unit of observation in this regression was a student (i) of a teacher (j) in a given grade (g), school (s), track (tr) and year (yr). The dependent variable ($s_{i,yr}$) was the student's standardized math or language arts test score taken in the spring of the school year. Students who did not take the spring test were excluded from the analysis (see discussion of attrition below). $Cert_j$ was a vector of indicators of the teacher's National Board certification status (passed, failed, or withdrawn) with non-applicants being the omitted category. The coefficients on these variables (λ_1) capture the difference in spring test scores between students of National Board applicants and non-applicants, and are the primary parameters of interest in that specification.

All specifications included fixed effects for school by grade by calendar track by year. In the experimental sample this amounted to including a fixed effect for each pair of teachers that was randomized, so that the coefficients were identified off of the within-pair variation (where teachers were randomized to class) rather than between pair. To ensure comparability, we used a similar identification strategy for the non-experimental sample, essentially comparing teacher impacts in the same school, grade, track and year. Standard errors in all analyses were clustered at the school-grade-calendar track-year level.

In the specification in Equation 1, we controlled for the student's baseline math, reading and language arts score ($S_{i, yr-1}$) from the previous spring testing (interacted with grade). Students missing the baseline score were imputed to the mean and dummies for missing test scores (interacted with grade) were included as controls. We also controlled for student characteristics ($X_{i, yr}$) including race/ethnicity (hispanic, white, black, other or missing), ever retained, title I, eligible for free lunch, homeless, migrant, gifted and talented, special education, english language development (level 1-5), and the means of these variables among all students in the class ($\bar{X}_{j, yr}^c$).

As a robustness check, we estimated models with and without the student and peer control variables ($S_{i, yr-1}, X_{i, yr}, \bar{X}_{j, yr}^c$). Omitting these control variables (but continuing to include the school-grade-track-year fixed effects) should not bias estimates of the difference between National Board applicants and non-applicants (λ_1) in the experimental sample, because class rosters were randomly assigned to teachers within each pair. Controlling for these baseline variables should only improve precision of the estimates in the experimental sample. In the non-experimental sample, omitting these control variables may lead to bias if National Board applicants and non-applicants are systematically assigned to students with different baseline characteristics. As a more direct test of whether National Board applicants and non-applicants are assigned to students with different baseline characteristics, we regress student baseline characteristics ($S_{i, yr-1}, X_{i, yr}$) on National Board status ($Cert_j$) and school-grade-track-year fixed effects. In the experimental sample we expect to find no significant difference between national

board applicants and non-applicants, while systematic sorting of students to National Board teachers may generate significant differences in the non-experimental sample.

We also tested whether National Board certification was related to teacher impacts on student achievement using a second specification closely related to equation 1. In the second specification, the dependent variable ($s_{i,yr} - s_{i,yr-1}$) was the change in the student's standardized score from the previous spring (with no imputing):

$$(2) \quad s_{i,yr} - s_{i,yr-1} = \lambda_1' Cert_j + \phi' X_{i,yr} + \gamma \bar{X}_{j,yr}^c + \delta'_{s,g,tr,yr} + \varepsilon'_{i,yr}$$

This specification controlled for student baseline achievement directly using test score gains, rather than including baseline test scores as a control (implicitly imposing a coefficient of one on the baseline test score). As in equation 1, this specification included fixed effects for school-grade-track-year, and we estimated specifications with and without controls for student and peer group characteristics ($X_{i,yr}, \bar{X}_{j,yr}^c$). Thus, the only difference between equation 2 and equation 1 was that equation 2 used test score gains rather than test score levels as the dependent variable, and did not control for the student's baseline test score. This method essentially imposes the assumption that the coefficient on baseline performance should be equal to one in equation (1) above. Although this should not matter for the experimental sample, it could have an impact on the non-experimental estimates, to the extent that measurement error in led us to understate the coefficient on prior performance.

Key Identifying Assumption: Within-School and Grade Variation versus Between

Even in a district as large as Los Angeles, there were few cases where a successful and an unsuccessful applicant were teaching in the same grade and subject.

As a result, without involuntarily moving teachers (or students) between schools and grades, it would not have been practical to use random assignment to compare NBPTS applicants in a “head-to-head” comparison. Rather, each NBPTS applicant is being compared to a comparison teacher in their school, grade and subject. To the extent that the comparison teachers assigned to more successful applicants were themselves more effective than the comparison teachers assigned to less successful applicants, then we may be understating the effects of NBPTS certification. We test this assumption by comparing the teacher-level impacts for the comparison teachers (estimated non-experimentally) assigned to high and low-scoring NBPTS applicants. If the comparison teachers for the more successful applicants were, indeed, more effective, we might expect to see some relationship between comparisons in different schools and the scaled scores of the NBPTS applicants from those schools.

Evaluating Other Threats to the Validity of the Experimental Estimates

There were two main potential threats to the validity of our estimates in the experimental sample. First, while class rosters were randomly assigned to teachers within each pair, not all of these students remained in the class with their assigned teacher and took spring tests in the following year. This could bias the experimental estimates if student attrition was large and differed systematically between students assigned to National Board applicants and non-applicants. To test for differential attrition, we estimated specifications identical to equation 1 (with and without the control variables) using as the dependent variable whether the student was missing their spring test score in

math or reading (separately) and whether the student switched to another teacher by the spring.

A second potential threat arose because the principal of each school had to agree to participate in the experiment (prior to randomization). If the National Board applicants or non-applicants in schools agreeing to participate were systematically different from applicants and non-applicants in other schools, then the experimental estimates would lack external validity. We used data from four years prior to the experiment (2000-2003) to test whether the teachers subsequently participating in the experiment had differed from other teachers with the same National Board status (passed, failed, withdrew, or non-applicant) in terms of their impact on student test scores in the years before the experiment. This test was based on specifications identical to equations 1 and 2.

Scaled Scores of National Board Applicants and Student Achievement

For all National Board applicants, we used information on their NBPTS scaled score—not just whether candidates achieved certification—to test whether the score is related to teacher impacts. To test the predictive value of the NBPTS score itself, we estimated regressions analogous to equations 1 and 2 in both the experimental and non-experimental samples:

$$(3) \quad s_{i,yr} = \lambda_2 \text{EverApplied}_j + \lambda_3 \text{NBScore}_j + \beta_g S_{i,yr-1} + \phi X_{i,yr} + \gamma \bar{X}_{j,yr}^c + \delta_{s,g,tr,yr} + \varepsilon_{i,yr}$$

$$(4) \quad s_{i,yr} - s_{i,yr-1} = \lambda'_2 \text{EverApplied}_j + \lambda'_3 \text{NBScore}_j + \phi' X_{i,yr} + \gamma' \bar{X}_{j,yr}^c + \delta'_{s,g,tr,yr} + \varepsilon'_{i,yr}$$

These equations replaced the indicators for National Board status ($Cert_j$) with an indicator for if the teacher had applied to the National Board ($EverApplied_j$) and, if so,

their NBPTS scaled score ($NBScore_j$). The NBPTS score was standardized to have mean zero and standard deviation one, and was set to zero for those who never applied to the National Board. Thus, the coefficient on the indicator for having ever applied (λ_2) represented the impact on student test scores of a National Board applicant with an average scaled score, relative to the impact of a non-applicant. The coefficient on the NBPTS score (λ_3) represented how much larger the impact was, relative to non-applicants, for a National Board applicant who scored one standard deviation higher on the NBPTS score. Applicants to the National Board who withdrew (and therefore did not have a scaled score) were dropped from the analysis.

Using similar specifications, we also tested the predictive value of each of 10 sub-scores which were aggregated by the NBPTS into a single scaled score. Lacking any student achievement data to validate against, the NBPTS established the weights arbitrarily for each of the components of the portfolio and assessment center exercises, without reference to student achievement impacts. We included each of the sub-scores separately in equations 3 and 4, and tested whether various subsets of the sub-scores were jointly significant. When all the sub-scores were included as separate regressors, their coefficients offer an estimate of the optimal weight that should be placed on each sub-score if the goal is to generate the best prediction of National Board applicants' impact on student test scores. We tested whether these estimated weights were significantly different from the weights imposed by the NBPTS scaled score.

Finally, we used specifications similar to equations 3 and 4 to evaluate the predictive power of the NBPTS scaled score against two alternatives. First, we used the coefficients on the sub-scores from the non-experimental sample to re-weight the sub-

scores and form our own score using these more optimal weights. We then compared the coefficients in equations 3 and 4 when we replaced the NBPTS-weighted score with the optimally weighted score. Second, we estimated the impact of the NBPTS scaled score while controlling for a non-experimental estimate of each teacher's "value-added" from prior years. We derived the value-added estimate for each teacher by estimating specifications analogous to equation 1 (excluding the indicators for National Board status) with data from 2000 to 2003, and then calculating the average residual for each teacher. Thus, a teacher with high value-added was a teacher whose student's had higher than expected spring test scores over these prior years. We standardized these teacher residuals to be mean zero and standard deviation one. For the academic years ending in the spring of 2004 and 2005, we then estimated equations 3 and 4 controlling for this additional measure of teacher value-added. These regressions estimated the marginal contribution of the NBPTS scaled score among those teachers with similar "value-added" estimates from prior years.

VII. Results

Before reporting estimates of impacts on student achievement, we first report evidence on the baseline characteristics of those assigned to various groups of applicants and non-applicants, as well as evidence on attrition and the likelihood of switching teachers by applicant status.

Baseline Characteristics

Table 4 reports differences in the baseline characteristics of students taught by National Board applicants—whether they achieved certification, did not achieve or were missing scaled scores from the National Board (many of these presumably withdrew from the National Board process before a final score was issued). The reported results included fixed effects for each permutation of school, grade and calendar track. The estimates in Table 4 report differences for each of the three groups (NBCT's, unsuccessful applicants and those with unknown scores) relative to non-applicants in the same grade. Each column in the table reports the finding for a different student characteristic: baseline math and language arts scores (in standard deviation units), gifted and talented participation, whether they were ever retained in class, whether they were special education students or participated in Title I or the Free/Reduced Price lunch program, race/ethnicity and English Language Development status. The top panel reports results for the *experimental sample*, while the bottom panel contains results for the *non-experimental sample*.

The bottom two rows report the p-values for two hypotheses: first, that the students assigned to all three groups of applicants (achievers, non-achievers and those withdrawing) are no different from those assigned to non-applicant teachers and, second, among the applicants, that those who achieved National Board certification had students who were no different than the applicants who did not achieve certification. For the experimental sample, the p-values of these hypotheses tests were all greater than .05, indicating that we could not reject the hypothesis that there were no differences in student baseline scores. However, on one characteristic—gifted and talented status—the

difference between National Board applicants and non-applicants the p-value for the first hypothesis test was .06. The fact that there was no statistically significant difference in baseline math or language arts scores and the other characteristics provides some reassurance that the random assignment process produced similar classes of students for each group of teacher. Moreover, the proportion of students participating in gifted and talented programs was not statistically different between National Board certified teachers and unsuccessful applicants (the second hypothesis described above).

However, as one might expect, the results were very different for the non-experimental sample. For many of the student characteristics reported—baseline math and language arts scores, gifted and talented status, special education status, Title I and Free/Reduced Price Lunch participation—we could reject the hypothesis that students assigned to National Board applicants were similar to students assigned to non-applicants. For instance, even *among those teaching in the same school, grade and calendar track*, National Board certified teachers and unsuccessful National Board applicants were assigned students with baseline test scores .15 and .12 standard deviations higher than students assigned to those who never applied to the National Board. In other words, National Board applicants are regularly assigned students who are stronger academically than those assigned to non-applicants within the same school. This underscores the importance of the experimental design.

Interestingly, although National Board applicants were assigned students that were statistically significantly different from non-applicants (the first hypothesis test above), successful and unsuccessful applicants seemed to be assigned similar students (the second hypothesis test reported in the table). (We could not reject the hypothesis of

no difference in student characteristics between “achievers” and “non-achievers” for all but one of the characteristics (Title I status)).

Attrition and Teacher Switching

Throughout our analysis, we will be studying the subsequent math and language arts performance of students initially assigned to National Board applicants and non-applicants—regardless of the classroom that they are placed at the end of the year. Our analysis focuses on estimating the effect of having been *assigned* a National Board applicant as one’s instructor at the beginning of the year (since that is the treatment that was randomly assigned) and not the impact of having *participated* in an applicant’s classroom for the whole school year.

At the end of the school year, we were able to observe math and language arts performance for 93.3 percent of the students initially assigned to one of the experimental sample classrooms. Moreover, 85 percent of those students assigned a given teacher at the beginning of the year was still assigned to the teacher at the end of the year. As a result, the impact of being assigned a National Board applicant will be similar to the impact of actually having been taught by a National Board applicant, since 85 percent of those assigned to a given teacher at the beginning of the year were still in that teachers’ classroom at the end of the year.⁹

However, in Table 5, we report differences in the proportion of students with missing math or reading scores or switching teachers for applicants and non-applicants. We could not reject the hypothesis that there was no difference in the likelihood of

⁹ The teacher switching variable is defined only for those students who had a valid teacher ID both at the beginning and the end of the year.

missing scores or switching teachers between the three groups of applicants and non-applicants in the experimental sample. In the non-experimental sample, there was a very small, but statistically significant difference in the proportion of students with missing math scores between the applicants and non-applicants. However, even in the non-experimental sample, there was no statistically significant difference in the proportion of students of applicants and non-applicants missing language arts scores or switching teachers.

Impact During the Experimental Period

Table 6 reports the estimated impacts on the California Standards Test during the experimental period (spring of 2004 and 2005) for the experimental sample of teachers (top panel) as well as for the non-experimental sample (bottom panel). The first four columns report results for math achievement, using end-of-year scores as well as gain scores as the dependent variable, with and without controlling for student and peer-level covariates. The last four columns report analogous estimates for language arts scores.

As reported in the first column for the experimental sample, students assigned to NBPTS-certified teachers outperformed those assigned to comparison teachers by .07 standard deviations, while those assigned on unsuccessful NBPTS applicants underperformed by -.11 standard deviations. Given the magnitude of the standard errors, neither of these differences is statistically significant. The difference between the two (between the NBCT impact and the unsuccessful applicant impact) is statistically significant only at the .14 level.

In the second column, we add student and classroom-level covariates. The resulting estimates are somewhat more precise. Although the difference between having

an NBCT and having a non-applicant teacher is not statistically significant (.046 standard deviations with a standard error of .049), students assigned to unsuccessful applicants under-perform similar students assigned to non-applicants by a statistically significant .17 standard deviations. As reported in the bottom of the panel, the difference between the certified teacher impact and the unsuccessful applicants is little different between columns (1) and (2)-- .18 (.07+.11) compared to .22 (.05+.17)—but the latter is statistically significant at the .01 level.

In columns (3) and (4), the dependent variable is the gain in math performance relative to the prior year—essentially imposing the assumption that the coefficient on prior year's score is equal to 1 for all grades. (The controls in column (4) include the same controls as in column (2), while excluding a student's prior year test score.) In gain scores, the pattern of impacts is similar to those in column (2)—with no statistically significant difference between NBCT's and non-applicants. Those assigned to unsuccessful applicants underperformed relative to those assigned to non-applicants.

The bottom panel reports results for the non-experimental sample. Given the lack of random assignment for this sample and the large differences in baseline performance reported in Table 4, we would expect large differences in column (1) before controlling for other actors. However, in column (2), when we add controls for student and classroom-level regressors, the estimated impacts are similar to those observed in the experimental sample, although somewhat smaller. While there was no statistically significant difference between those assigned to NBCT's and non-applicants, those assigned to unsuccessful applicants *underperformed* by .07 standard deviations relative to those in the classrooms of non-applicants. The difference in the two impacts was

statistically significant at modest levels ($p\text{-value}=.067$). The results in columns (3) and (4) are similar: using gain scores, students assigned to unsuccessful applicants underperformed by .05 and .07 standard deviations and the difference in impacts between certified teachers and unsuccessful applicants significant at the .07 and .08 levels respectively.

When language arts achievement is the outcome, we continue to find differing impacts between NBCT's and unsuccessful applicants relative to non-applicants in the experimental sample—from .18 to .25 standard deviations. These differences are statistically significant in columns 6 through 8 (which control for baseline performance either by adding a regressor or using a gain score). For the non-experimental sample, there is no estimated impact on language arts achievement.

Impacts During the Pre-Experimental Period

Although they are qualitatively similar, the estimated impact of having an NBCT rather than an unsuccessful applicant for the experimental sample is two to three times larger than for the non-experimental sample. One possible explanation is that the 99 pairs of teachers chosen for the experiment-- either the NBPTS applicants or the comparison teachers-- could be non-representative. To test this hypothesis, we generate non-experimental estimates of the impacts during the *pre-experimental period*—2000-03—for those teachers subsequently included in the experimental and non-experimental samples.

The results of this analysis are reported in Table 7. We report the results from three specifications—no controls (except for school by grade by calendar track by year

fixed effects), the full set of student and peer controls and student fixed effects with peer controls. We do so using math and language arts as the outcome. We report the p-values for a series of hypothesis tests at the bottom of Table 7. In the first column with no controls, we find that those students assigned to the comparison teachers in the experimental sample performed slightly *better* than the students assigned to other non-applicants not chosen to be part of the experiment (.09 standard deviations with a p-value=.07). However, we could not reject the hypothesis that the students assigned to the subset of NBCT's or the unsuccessful applicants chosen for the experiment performed the *same as* students assigned to the NBCT's or unsuccessful applicants that were not chosen (p-value=.6567).

After including the full set of student-level and classroom peer controls in column 2, our estimates for the non-experimental sample in the pre-experiment years are very similar to those in the experiment years—with a .099 (.048+.051) difference in the impact of having been assigned an NBCT versus an unsuccessful applicant. We could not reject that hypothesis the experimental comparison group of non-applicant teachers had the same impact as the non-experimental comparison group (p-value=.8157). Moreover, we could not reject the hypothesis that the experimental sample of NBCT's and unsuccessful applicants had the same impact as the non-experimental sample during these years (p-value=.4737.)

In the third column, we include fixed effects for permutations of students and the schools they attended. Again, we find similar estimates to those reported for the non-experimental sample during the years of the experiment-- .104 (.039+.065) standard deviation difference in the impact of NBCT's and unsuccessful applicants. Moreover, we

could not reject the hypotheses that the experimental controls were no different from other non-applicants (p-value=.6572) nor that the NBCT's and unsuccessful applicants in the experimental sample had the same impact as the non-experimental sample during the pre-experiment years (p-value=.6572).

With language arts as the outcome, we also fail to find evidence that the set of NBCT's, unsuccessful NBPTS applicants or comparison teachers chosen for the experimental sample were having differing impacts in the pre-experiment years than those in the non-experimental sample.

The results reported in Table 7, therefore, provide little reason to believe that the experimental sample was “cherry-picked” in a way which would have led us to find larger effects of NBPTS certification. Although we continue to look into potential explanations of the difference between the experimental and non-experimental results, it is worthwhile noting that a similar pattern has been observed in the evaluation of the impact of Teach for America corps members.¹⁰ The experimental evaluation of Teach For America by Decker, Mayer and Glazerman (2004) reported impacts of .15 student-level standard deviations in math. The non-experimental evaluations of that program, such as by Kane, Rockoff and Staiger (2006) have reported considerably smaller impacts (.02 student-level standard deviations in math and no statistically significant impact on reading.)

Scaled Score vs. Certification Status

¹⁰ We will seek to obtain teacher absence data before, during and after the experiment for NBPTS applicants the comparison teachers included in the evaluation.

A dichotomous measure—such as whether one is certified or not-- simply does not contain as much information as the scaled score itself. As a result, the traditional approach of comparing the impacts of the NBCT's with the unsuccessful applicants conflates any information in the scaled score about with the distribution of scaled scores above and below the cut-off.

Figure 1 displays the distribution of NBPTS scaled scores for applicants working in the Los Angeles Unified School district each year from 1999 through 2004. The vertical line is drawn at the cut-off for National Board certification of 275. The distribution of scaled scores—as in the national population—is centered near the cut-off for certification. This is reflected in the fact that approximately half of those taking the exam in any given year achieve certification.

Just changing the cut-off can have a large effect on the difference in mean scaled scores for those above and below the cut-off. To illustrate this point, we re-calculated the difference in mean scaled scores for “achievers” and “non-achievers” for all cut-offs from 200 through 350. As reported in Figure 2, the difference between those above and below the cut-off is close to its minimum at the cut-off of 275. The difference in mean scaled scores at that point is 53 scaled score points—about 1.5 standard deviations. This is a result of the NBPTS's decision to place the cut-off near the mean of the scaled score distribution. By ensuring that roughly 50 percent of applicants in a given year achieve certification, the board ensured that the difference in mean scaled score between those passing and those failing was at its minimum.

While districts are interested in the specific question of whether NBCT's outperform unsuccessful applicants or whether NBCT's outperform non-applicant

(because that is the way their bonus policies are designed), they should also be interested in how much information the National Board process generates—that is, how much information is there in the scaled score, or how much information could there be in the scaled score if the sub-scores were weighted differently. The remainder of the paper focuses on those two questions.

First Score or Maximum Score?

For a given NBPTS applicant, we might have several different test scores, if the applicant were to retake the exam several times. To test the validity of the two measures, we first calculated the non-experimental value-added estimates for each NBPTS applicant during the pre-experimental period, 2000 through 2003. To generate these estimates, we used a two-step process. First, we first estimated teacher effects separately by year from spring 2000 through 2003, conditioning on student test scores from the previous spring as well as demographic and program participation indicators. Second, we took the mean of the residuals, after accounting for fixed effects by school, grade, calendar track and year as well as classroom-level covariates. We did this separately for math and language arts, although we will be focusing on the math results. In addition, because we use these point estimates later in the paper to validate against the experimental results, we dropped any student from the pre-experimental sample, who was included in the experimental sample.

We then calculated the running mean of the pre-experimental estimate of value-added by scaled score, taking 30 observations to the right and left of the current value of the scaled score (a running mean of a total of 60 observations). We repeated the exercise

for the *maximum* scaled score as well as the *first* scaled score observed for each NBPTS applicant, and reported the 95 percent confidence interval for each running mean.

As reported in Figure 3, the relationship between pre-experimental value-added and the first scaled score received by an applicant is upward sloping and fairly linear between 200 and 325 (roughly the 5th and 95th percentiles). Figure 3 also reports the *maximum* scaled score an applicant received. Figure 3 implies that there may be a dip in mean value-added immediately after the cut-score for passage. There's a large number of applicants with scores just below the cut-off who, upon retaking the exam, score just above the cut-off for passage. As a result, there is a small dip in performance just above the passing cut-off. As a result, in the remainder of the paper we will focus on testing the relationship between an applicants' first scaled score, rather than their ultimate scaled score.

Validating the Scaling of the NBPTS Sub-scores in the Non-Experimental Sample

The weights attached to each component of a teacher's portfolio and assessment center essays were chosen based on the board's professional judgment. However, this is a difficult assignment to tackle based on intuition alone. At the time such weights were established in the early Nineties, few states possessed longitudinal data for students and teachers. In this section, we use the data for the non-experimental sample to estimate a weight on each of the ten subcomponents separately. We then test the linear constraint implied by the relative weighting used by the NBPTS—under both the old and the new system.

To do so, we used the non-experimental sample to estimate the following specification:

$$s_{i,yr} = \beta_0 + \beta_1 Applied_j + Applied_j * \left(\sum_{k=1}^{10} \lambda_k^{old} NBSubScore_{jk}^{old} + \sum_{k=1}^{10} \lambda_k^{new} NBSubScore_{jk}^{new} \right) + \beta_g S_{i,yr-1} + \phi X_{i,yr} + \gamma \bar{X}_{j,yr}^c + \delta_{s,g,tr,yr} + \varepsilon_{i,yr}$$

Where $Applied_j$ is equal to 1 for NBPTS applicants, $NBSubScore_k^{old}$ and $NBSubScore_{k1}^{new}$ is equal to an applicant's score on the k^{th} component of the NBPTS assessment under the old or the new system (demeaned so that they have a mean of zero and for non-applicants is equal to zero), $s_{i,yr-1}$ is a vector of a student's math, reading and language arts score from the prior spring, $X_{i,yr}$ is the student's demographic and program participation characteristics, $\bar{X}_{j,yr}$ are the mean characteristics of the students in the class and $\delta_{i,g,tr,yr}$ are fixed effects for the permutation of school by grade by calendar track and year. In the specification above, β_1 measures the difference between applicants (evaluated at the mean on each sub-scores) and non-applicants and the coefficients on the sub-scores, λ^{new} and λ^{old} , measure the difference in impact of a National Board applicant relative to the comparison non-applicant per 1 unit change in the sub-score.

Table 8 reports a series of hypothesis tests involving the coefficients on the sub-scores, when math and language arts scores are used as the outcome, $s_{i,yr}$. When predicting student achievement in both math and language arts, would could reject the hypothesis that the coefficients on all the sub-scores are equal to zero. However, it was

difficult to pin down which of the components were superfluous. When predicting math achievement, the only components for which we could not reject the hypothesis of no impact was the teacher commentary on student work. With math achievement as the dependent variable, we could reject the hypothesis that the weights attached to the video scores were zero. The same was true for the assessment center and documented accomplishment exercises. However, the data seem not to prefer the relative weights chosen by the National Board. We could strongly reject the hypothesis that both the relative weights equaled those in either the old or the new indices (p-value of .0133 and .0069) respectively.

In predicting language arts achievement, the results were less clear. While we could reject the hypothesis that all the components should be weighted with a zero weight, we *could not reject* the hypothesis that any of the categories of scores taken alone—videos, student work, documented accomplishments and assessment center exercise—were equal to zero.

Combining the Predictive Power of the NBPTS Scaled Score and Prior Value Added

As noted above, we are ultimately interested in learning how much the current scaled score contributes to predicting teacher impacts. We are also interested in learning whether that predictive power could easily be improved simply by re-weighting the various components of the index. Finally, we are interested in learning whether either of these measures *add to* other pieces of information—such as estimates of prior value-added in prior years—in identifying effective teachers. To pursue these questions, we first used the weights implied by this validation exercise-- λ^{new} and λ^{old} -- to calculate a

new “imputed scaled score” for each National Board applicant. We also calculated the mean teacher effect for each teacher in the pre-experimental period, 2000 through 2002. In Table 9, we evaluate the predictive power of the National Board’s scaled score against the “imputed scaled score” as well as the prior non-experimental estimate of value-added.

We report the results of three specifications similar in form to those in Table 8. Because the scaled scores and prior value-added measures have been standardized to have a mean of zero (as well as a standard deviation of one), the coefficient on the indicator for NBPTS application identifies the difference between the applicant with the mean scaled score and the mean non-applicant. Across all specifications, we could not reject the hypothesis that the mean applicant to the National Board had a similar impact as the mean applicant. In other words, the National Board application process in Los Angeles is drawing roughly the mean teacher.

The first column includes the standardized version of the actual NBPTS score. A one standard deviation difference in performance on the scaled score is associated with a .11 standard deviation difference in impact on student performance in the experimental sample and .06 in the non-experimental sample. Both results are consistent with our estimates of the binary effect of certification in a similar specification in Table 6. With a certification cut-off at the mean, the difference between those with scores above and below the cut-off is roughly 1.5 standard deviations on the scaled score, which when multiplied by the coefficient in columns (1) and (4) would imply a .09 standard deviation difference in mean performance between achievers and non-achievers in the non-experimental sample ($1.5 \times .06 = .09$) and .17 for the experimental sample ($1.5 \times .11 = .17$).

The second column continues to control for the standardized version of the NB scaled score, but adds the measure of value-added that we calculated during the pre-experimental period of 2000 through 2002. This prior non-experimental estimate of value-added has a statistically significant coefficient of .19 in the non-experimental sample and .20 in the experimental sample. This may not be surprising in the non-experimental sample since any biases which led us to overstate or understate a given teacher's impact in 2000 through 2002 may carry over into the later period. Other studies have confirmed that there is a correlation in non-experimental value-added estimates over time. However, prior value-added has a similar effect even within the randomized pairs. Each one standard deviation difference in prior value-added—estimated non-experimentally—is associated with a .20 standard difference in performance between pairs of teachers randomly assigned within the experiment.

With the inclusion of the prior estimate of value-added, the coefficient on the National Board standardized score in column (2) is considerably smaller and no longer statistically significant for either the experimental or non-experimental samples. Although the National Board's scaled score contains information that is helpful in predicting a teacher's effectiveness, that information seems to be contained within the prior value-added estimate as well. At least when predicting math or reading achievement on the California Standards Test, there is no additional information provided by the NBPTS scaled score once prior estimates of value-added are included as a covariate.

In the third column, we replace the actual NBPTS scaled score with the “imputed” NBPTS scaled score, using the weights on the sub-scores estimated in Table 7. (Recall

that these were imputed with only the non-experimental sample and not the experimental sample.) The coefficient on the imputed scaled score implies that a 1 standard deviation difference in that score is associated with a .08 difference in impact for the non-experimental sample (with a p-value less than .01). The coefficient is of a similar magnitude in the experimental sample .07, but is only marginally statistically significant (p-value=.07).

The results using language arts as the outcome are roughly similar, with the predictive power of the prior estimate of value-added particularly strong in the experimental sample. The predictive power of the imputed scaled score is statistically significant in the non-experimental sample, but not in the experimental sample.

Testing the Between-School Comparability of Comparison Teachers

For purely practical reasons, the experimental design focused on within-school comparisons. (It is hard to imagine a school district ever agreeing to have experienced teachers or students randomly assigned across schools!) Our non-experimental estimates mimicked that design by including school, grade and calendar track fixed effects.¹¹ However, a very strong assumption implicit in that design is that quality of the *comparison* teachers working in the same school, grade and track is unrelated to the National Board applicant's scaled score. In other words, we are assuming that comparison teachers working in schools with the highest-scoring National Board applicants are similar to the comparison teachers working with the lowest-scoring

¹¹ There may be other reasons to include school fixed effects, such as to control for unmeasured differences in student background characteristics.

applicants. If the comparison teachers working with NBPTS-certified are better than average, we may be understating the effect of the scaled score.

Although the random assignment only ensured valid comparisons within school, we used the non-experimental methods to evaluate the relationship between scaled scores and student achievement across schools and grades. Limiting the sample to students taught by National Board applicants, we estimated the following specification:

$$S_{it} = \lambda_1 ScaledScore_j + \beta_{g1} S_{i,t-1} + \phi_1 X_{it} + \varepsilon_{it}$$

where j subscripts the teacher, g the grade, i the student and t the year. X represents a vector of student level characteristics (the same used in previous analyses) and $S_{i,t-1}$ represents student's scores from the previous spring. In other words, among those taught by NB applicants, even if they are in different schools, is student-achievement any higher after controlling for students' baseline achievement.

Limiting the sample to students assigned to comparison teachers working in the same grade, school and calendar track as NB applicants (that is, excluding students taught by NB applicants and those where there is no NB applicant in the school/grade), we estimated the following specification:

$$S_{it} = \lambda_2 ScaledScore' + \beta_{g2} S_{i,t-1} + \phi_2 X_{it} + \varepsilon_{it}$$

where the variable ScaledScore' measures the mean scaled score of the NB applicants in that comparison teachers' school, grade and calendar track. If the effectiveness of the comparison teachers matched with "achievers" is higher than the effectiveness of those assigned to "non-achievers" the coefficient λ_2 would be positive.

Table 10 reports the results of both specifications. Among those assigned to National Board applicants, the students assigned to teachers with higher scaled scores outperformed similar students assigned to those with lower scaled scores. Within the non-experimental sample, a one-standard deviation difference in scaled scores was associated with .085 and .056 standard deviation increase in math and language arts performance respectively, holding constant baseline test scores and student demographics. The point estimates are positive in the experimental sample, but they were not statistically significant.

Table 10 also suggests that there was no relationship between the comparison teachers' effectiveness and the scaled score of the NBPTS applicant they were matched with. Among those assigned to comparison teachers, there is no relationship between the scaled score of the NBPTS applicants in their grade/school/calendar track and their own effectiveness. In other words, we find no evidence that the comparison teachers matched with NBPTS certified teachers were any more effective than those matched with unsuccessful applicants.

Regressions using Pair-Level Differences in Means from the Experimental Sample

To further probe the robustness of the above results, we estimated a number of simple bivariate regressions using pair-level differences in means from the experimental sample. The results are reported in Table 11. The first column reports coefficients from two separate regressions. The dependent variable in both is the *difference* in mean baseline math performance between students assigned to the NBPTS applicant and students assigned to the comparison teacher. There is only one observation per pair. The

regressors are the difference in prior value-added estimated non-experimentally during the years 2000 through 2003. (For these results, we have left the prior value-added estimates in the same units as the dependent variable—teacher impact on student-level achievement.)

The first row reports the bivariate regression coefficient on the difference in prior value-added and the second row reports the regression coefficient from regressing the difference in baseline scores on the candidate's NBPTS score. Reflective of the random assignment of classrooms within pairs, neither coefficient is statistically significant.

The second column uses the difference in student achievement at the end of the year. A one standard deviation difference in teachers' pre-experimental estimates of value-added in math (estimated non-experimentally) was associated with a .2 standard-deviation difference in math performance at the end of the year. The data underlying that regression are plotted in Figure 3. As Todd and Wolpin (2003) have reminded us, there are a number of strong assumptions implicit in the conventional non-experimental value-added specification—probably the most important of which is that prior year test performance is a sufficient statistic for all prior educational inputs. As far as we know, this is the first effort to validate non-experimental value-added estimates in an experiment.

The third column uses the difference in the mean gain in student performance (relative to the baseline) within each pair of teachers as the dependent variable. The coefficient on prior value-added is .51, which is statistically different from zero. As we had reported in Table 6, the results in the second row of Table 11 imply that the NBPTS applicants' with larger scaled scores had somewhat larger gains relative to their

comparison teachers than those with lower scaled scores. (While significant in gains, this relationship is not significant in levels.) In the final column, we used the prior non-experimental value-added estimate for each National Board applicant as the dependent variable. Consistent with the results in Table 7, the NBPTS applicant's scaled score is related to their prior estimate of value-added (coefficient equal to .0014 with a p-value of .05), although the point estimate is somewhat smaller than that observed in the experiment.

VI. Conclusion

The NBPTS scoring process captures information that is helpful in identifying effective teachers. However, that information is not being used efficiently. The results in this paper suggest a number of potential improvements—creating multiple levels of performance rather than a single binary measure, recalculating the scaled score using re-weighted sub-scores and, potentially, preventing applicants from retaking the exam multiple times. Moreover, our results suggest that the student work component may provide relatively little information and that some subsets of the subscores—such as video scores—capture much of the information.

Our results also suggest that prior non-experimental estimates of “value-added” are helpful in predicting differences in student outcomes in an experimental setting. Even among pairs of teachers for whom classroom rosters were randomly assigned, those students assigned to teachers with high prior value-added estimates significantly outperformed those with low value-added scores. Given the growing reliance on “value-

added” techniques in education research, this is a fundamentally important finding that we are pursuing further in related work.

Practice-based approaches to assessing teacher performance, such as the NBPTS application process, have typically been portrayed as being at odds with the value-added approach. This is an unfortunate historical accident, driven more by the ideological predispositions of their respective supporters, rather than any substantive reason. Our results imply that the *combination of both* the NBPTS scores and the prior value-added estimates could be helpful in identifying those teachers most likely to produce exemplary student gains. In those grades and subjects where value-added assessments are practical, the NBPTS should consider incorporating a value-added measure as an additional sub-score contributing to their scaled scores.

The ultimate value of any signal of effective teaching—such as NBPTS certification—depends not only on its usefulness in predicting future performance, but also on the nature of policy response. If the information leads to no change in behavior—has no effect on who ends up in a classroom, what they do, or whom they are assigned to teach—it has no value, regardless of its predictive power. The average applicant to the National Board for Professional Teaching Standards has been in the classroom for 13 years. The NBPTS process may yield information of useful predictive power, but it may come too late in a teacher’s career to be of much use. A district could provide bonuses to such teachers, but if teachers have already demonstrated a commitment to a career in teaching, such bonuses could end up having little impact on retention. (Teachers seem unable to anticipate with much accuracy their chances of success given that roughly half of those taking the exam fail and the mean value-added of applicants is similar to those of

non-applicants.) Such bonuses may be well-deserved and may fulfill our notions of fairness, but student achievement will not be affected unless they lead to increases in retention among the most effective teachers (and if those bonuses come only after teachers have been in a district for 13 years, that seems unlikely).

Therefore, it is worth asking whether the value of the information provided by NBPTS might be improved simply by generating the information *earlier* in teacher's careers. Obviously, it would need to be demonstrated that the process is equally able to discern effective teaching during the first three years. However, it is during that initial two to three years of a teacher's career that teachers are exploring their commitment to teaching and collective bargaining agreements allow districts to discontinue the contracts of ineffective teachers.

References:

- Bommer, W.H., J.L. Johnson, G.A. Rich, P.M. Podsakoff, S.B. Mackenzie. 1995. "On the Interchangeability of Objective and Subjective Measures of Employee Performance: A Meta-Analysis." *Personnel Psychology* 48: 587-605.
- Cavalluzzo, Linda C. 2004. *Is National Board Certification An Effective Signal of Teacher Quality?* Alexandria, Virginia: The CNA Corporation.
- Clotfelter, Charles T., Helen Ladd and Jacob Vigdor "How and Why Do Teacher Credentials Matter for Student Achievement?" NBER Working Paper 12828, January 2007.
- Daley, Glenn and Rosa Valdés. 2006. *Value Added Analysis and Classroom Observation as Measures of Teacher Performance: A Preliminary Report.* Los Angeles Unified School District.
- Decker, Paul T., Daniel P. Mayer and Steven Glazerman. (2004) "The Effects of Teach For America on Students: Findings from a National Evaluation," *Mathematica Policy Research Report No. 8792-750*, June 9, 2004.
- Gallagher, H. Alix. 2004. "Vaughn Elementary's Innovative Teacher Evaluation System: Are Teacher Evaluation Scores Related to Growth in Student Achievement?" *Peabody Journal of Education* 79(4): 79-107.
- Goldhaber, Dan, and Emily Anthony. (forthcoming). "Can Teacher Quality Be Effectively Assessed? National Board Certification as a Signal of Effective Teaching." *Review of Economics and Statistics*.
- Goldhaber, Dan, David Perry, Emily Anthony. 2004. "The National Board for Professional Teaching Standards (NBPTS) Process: Who Applies and What Factors Are Associated with NBPTS Certification?," *Educational Evaluation and Policy Analysis* 26(4):259-280.
- Harris, Douglas N. and Tim R. Sass. 2007. *The Effects of NBPTS-Certified Teachers on Student Achievement.* Technical Report. NBPTS
- Heneman, Robert L. 1986. "The Relationship between Supervisory Ratings and Results-Oriented Measures of Performance." *Personnel Psychology* 1986: 811-826.
- Jacob, Brian A. and Lars Lefgren. 2005. "Principals as Agents: Subjective Performance Measurement in Education." Faculty Research Working Paper Series, RWP05-040. John F. Kennedy School of Government.
- Kane, Thomas J., Jonah Rockoff and Douglas Staiger, "What Does Certification Tell Us about Teacher Effectiveness?: Evidence from New York City" *NBER Working Paper No. 12155*, April 2006.

- Kimball, Steven M., Brad White, and Anthony T. Milanowski. 2004. "Examining the Relationship between Teacher Evaluation and Student Assessment Results in Washoe County." *Peabody Journal of Education* 79(4): 54-78.
- LAUSD Office of Communications. 2006. "Fingertip Facts 2005-06." http://notebook.lausd.net/pls/ptl/docs/PAGE/CA_LAUSD/LAUSDNET/OFFICE_S/COMMUNICATIONS/COMMUNICATIONS_FACTS/FACTSHEET_ENGLISH%20FINGERTIP%20FACTS%2005-06.PDF.
- LaLonde, Robert "Evaluating the Econometric Evaluations of Training Programs with Experimental Data" *American Economic Review* (1986) Vol. 76, pp. 604-620.
- McCaffrey, Daniel F., J. R. Lockwood, Daniel M. Koretz, and Laura S. Hamilton. 2003. *Evaluating Value-Added Models for Teacher Accountability*. Santa Monica, CA: RAND Corporation.
- Milanowski, Anthony. 2004. "The Relationship between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati." *Peabody Journal of Education* 79(4): 33-53.
- National Board of Professional Teaching Standards. 2006a. "2006 guide to National Board Certification." NBPTS.
- National Board of Professional Teaching Standards. 2006b. *Handbook on National Board Certification*. NBPTS.
- National Board of Professional Teaching Standards. 2006c. National Board Certification® Assessment Center Descriptions early Childhood/Generalist (Age Range: 3-8). NBPTS.
- National Board of Professional Teaching Standards. 2006d. *Portfolio Instructions – Middle Childhood Generalist*. NBPTS.
- National Board of Professional Teaching Standards. 2006e. National Board Certification® Assessment Center Descriptions middle Childhood/Generalist (Age Range: 7-12). NBPTS.
- Podgursky, Michael 2001. "Defrocking the National Board: Board: Will the Imprimatur of 'Board Certification' Professionalize Teaching?" *Education Matters* 1(2):79-82.
- Sanders, William L., James J. Ashton, and S. Paul Wright. 2005. Comparison of the Effects of NBPTS Certified Teachers with Other Teachers on the Rate of Student Academic Progress. Technical Report. SAS Institute.
- Todd, Petra E. and Kenneth I. Wolpin "On the Specification and Estimation of the Production Function for Cognitive Achievement" *The Economic Journal* (2003) Vol. 113, pp. F3-F33.

Vandevoort, Leslie G., Audrey Amrein-Beardsley, and David C. Berliner. 2004.
“National Board Certified Teachers and Their Students' Achievement.” *Education
Policy Analysis Archives* 12(46).

Table 1. Components of the NBPTS Application

| Entries | Name/Subject | Description of entry/exercise |
|------------------------------|---|---|
| Portfolio Entry 1 | Writing: Thinking through the Process | Written commentary on student work responding to two prompts generated by the candidate |
| Portfolio Entry 2 | Building a Classroom Community through Social Studies | Videotape of lesson with instructional materials used and written commentary on the lesson |
| Portfolio Entry 3 | Integrating Mathematics with Science | Videotape of lesson with instructional materials used and written commentary on the lesson |
| Portfolio Entry 4 | Documented Accomplishments: Contributions to Student Learning | Descriptions and documentation of Ability to partner with students, parents, and the learning community of the school to promote students' academic achievement |
| Assessment Center Exercise 1 | Supporting Reading Skills | Identify and interpret student errors through analyzing a transcript of a student's oral reading of a passage. Provide and justify appropriate strategies to address the identified needs of the student. |
| Assessment Center Exercise 2 | Analyzing Student Work | Identify and interpret mathematical misconceptions in sample student work. Provide and justify appropriate strategies to address the identified needs of the student. |
| Assessment Center Exercise 3 | Knowledge of Science | Teachers asked to respond to student inquiry in a way that demonstrates their understanding and ability to teach fundamental concepts and principles in science. |
| Assessment Center Exercise 4 | Social Studies | Teachers asked to interpret cause-and-effect relationship based on a given graphic image. Also asked to describe activity that would develop student understanding of this real world relationship. |
| Assessment Center Exercise 5 | Understanding Health | Teachers asked to identify health needs of a sample student and what steps or resources |

| | | |
|---------------------------------|----------------------|---|
| | | should be used to meet the needs of the student. |
| Assessment Center Exercise 6 | Integrating the Arts | Teachers asked to describe an arts-focused learning experience that would help students understand an identified concept in another discipline. Teachers also asked to explain how this will deepen the student's appreciation of the arts. |

Source: National Board of Professional Teaching Standards (2006d, 2006e)

Table 2. Weights Used to Calculate Scaled Scores

| Type of Entry/ Exercise | Before 2002 | | 2002 and After | |
|------------------------------------|--------------------|--------------------------------------|-----------------------|--------------------------------------|
| | Items | Total weight of all items | Items | Total weight of all items |
| Video | 2 | 24 | 2 | 32 |
| Commentary on student work | 2 | 24 | 1 | 16 |
| Documented accomplishments | 2 | 12 | 1 | 12 |
| Assessment Center exercises | 4 | 40 | 6 | 40 |
| TOTAL | 10 | 100 | 10 | 100 |

Table 3: Summary of Prior Research on NBPTS Certified Teachers and Students' Math Achievement

| Study | Pass v. fail effect size | Significant at 5% level? | Definition of "fail" | Pass v. "other" effect size | Significant at 5% level? | Definition of 'other' |
|---|--------------------------|--------------------------|--|-----------------------------|--------------------------|--|
| Goldhaber and Anthony (2005) | .09 | Yes | Applies but does not pass; completeness of application not known | .05 | Yes | Ever passed v. never applied |
| Cavaluzzo (2004) | .1 | Yes | Failed/withdraw; pending separate category | .07 | Yes | Certified v. not involved |
| McColskey et al. (2005) | NA | NA | NA | .07 | No | Board certified v. non-board certified |
| Clotfelter, Ladd, et al. (2006) | NA | NA | NA | .02-.03 | Yes | Board certified v. non-board certified |
| Harris and Sass (2006) | NA | NA | NA | -.01 | No | Ever Certified v. never certified |
| Sanders et al. (Model 2, Grades 4 and 5 pooled) | .07 | No | Unclear whether 'fail' includes incomplete applications and/or those withdrawing | .04 | No | Certified v. no involvement |

Table 4. Difference in Baseline Student Characteristics for Those Taught by NBPTS Certified Teachers and Unsuccessful Applicants

| Baseline Academic Performance | | | | | Baseline Demographics | | | | Baseline English Language Status | | | |
|-------------------------------|----------------|---------------------|---------------|-------------------|-----------------------|-------|---------|------------|----------------------------------|-----------|-------------|------------|
| Math Score | Language Score | Gifted and Talented | Ever Retained | Special Education | Hispanic | Black | Title I | Free Lunch | Level One | Level Two | Level Three | Level Four |

A. Experimental Sample

| | | | | | | | | | | | | | |
|-----------------|-------------------|-------------------|-------------------|-------------------|-------------------|------------------|-------------------|------------------|------------------|------------------|-------------------|-------------------|-------------------|
| National Board: | | | | | | | | | | | | | |
| Achieved | 0.059 (0.077) | 0.059 (0.089) | 0.036 (0.021) | 0.004 (0.014) | -0.007 (0.009) | 0.004 (0.013) | -0.008 (0.007) | 0.011 (0.012) | 0.002 (0.016) | 0.009 (0.012) | 0.03 (0.027) | -0.021 (0.026) | -0.004 (0.019) |
| Withdrew | -0.044 (0.141) | -0.099 (0.159) | -0.042 (0.030) | -0.02 (0.021) | -0.025 (0.024) | 0.047 (0.043) | -0.048 (0.041) | 0.02 (0.019) | 0.025 (0.021) | 0.045 (0.027) | 0.03 (0.027) | -0.02 (0.050) | -0.008 (0.023) |
| Did Not Achiev | 0.074 (0.055) | 0.18 (0.127) | 0.007 (0.011) | -0.006 (0.030) | -0.028 (0.025) | 0.038 (0.050) | -0.027 (0.024) | 0.024 (0.027) | 0.036 (0.028) | 0.013 (0.012) | -0.047 (0.046) | 0.09 (0.063) | -0.046 (0.033) |
| # Observations | 2,321 | 2,323 | 3,873 | 3,873 | 3,873 | 3,873 | 3,873 | 3,873 | 3,873 | 3,873 | 3,873 | 3,873 | 3,873 |
| p-values: | | | | | | | | | | | | | |
| Jointly=0 | 0.49 | 0.43 | 0.16 | 0.79 | 0.43 | 0.60 | 0.27 | 0.43 | 0.37 | 0.21 | 0.34 | 0.42 | 0.56 |
| Passed=Failed | 0.88 | 0.44 | 0.22 | 0.76 | 0.44 | 0.50 | 0.46 | 0.67 | 0.29 | 0.79 | 0.16 | 0.11 | 0.27 |

B. Non-Experimental Sample

| | | | | | | | | | | | | | |
|-----------------|-------------------|-------------------|-------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| National Board: | | | | | | | | | | | | | |
| Achieved | 0.1479 (0.028) | 0.1411 (0.029) | 0.0735 (0.011) | -0.0087 (0.005) | -0.0081 (0.004) | -0.0145 (0.008) | 0.0019 (0.005) | -0.0306 (0.008) | -0.0167 (0.006) | -0.0082 (0.004) | -0.0121 (0.009) | -0.0311 (0.012) | -0.0095 (0.010) |
| Withdrew | 0.241 (0.069) | 0.2376 (0.074) | 0.0861 (0.031) | 0.001 (0.012) | -0.0199 (0.009) | 0.0031 (0.024) | -0.0207 (0.017) | -0.0121 (0.024) | -0.0249 (0.015) | -0.0168 (0.010) | -0.0409 (0.026) | -0.0207 (0.026) | 0.0194 (0.020) |
| Did Not Achiev | 0.1164 (0.050) | 0.1047 (0.052) | 0.0499 (0.017) | -0.0211 (0.007) | -0.0126 (0.006) | -0.0206 (0.014) | 0.009 (0.009) | 0.0015 (0.009) | -0.0268 (0.010) | 0.0016 (0.006) | 0.0024 (0.014) | -0.0561 (0.019) | -0.0012 (0.014) |
| # Observations | 251,854 | 251,560 | 272,062 | 272,062 | 272,062 | 272,062 | 272,062 | 272,062 | 272,062 | 272,062 | 272,062 | 272,062 | 272,062 |
| p-values: | | | | | | | | | | | | | |
| Jointly=0 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.14 | 0.42 | 0.00 | 0.00 | 0.04 | 0.23 | 0.00 | 0.52 |
| Passed=Failed | 0.56 | 0.52 | 0.23 | 0.15 | 0.54 | 0.68 | 0.46 | 0.01 | 0.35 | 0.11 | 0.35 | 0.24 | 0.61 |

Note: All estimates control for school-by-year-by-grade-by-calendar track fixed effects. Dependent variables are baseline student characteristics from the prior school year. Experimental estimates include only pairs of teachers who were randomized to classrooms, while non-experimental estimates include all other teachers teaching in school-grade-years with a National Board applicant. Standard errors (in parentheses) allow for clustering at the school-grade-year level.

Table 5. Attrition and Teacher Switching

| Without Control Variables | | | With Control Variables | | |
|---------------------------|------------------------|------------------|------------------------|------------------------|------------------|
| Missing Math Score | Missing Language Score | Switched Teacher | Missing Math Score | Missing Language Score | Switched Teacher |

A. Experimental Sample

National Board:

Achieved -0.014 -0.017 0.009 -0.006 -0.009 0.019
 (0.012) (0.012) (0.038) (0.006) (0.006) (0.039)

Withdrew -0.012 -0.015 -0.164 0.001 -0.004 -0.157
 (0.017) (0.017) (0.087) (0.010) (0.010) (0.095)

Did Not Achieve -0.039 -0.035 0.015 -0.017 -0.014 0.01
 (0.033) (0.032) (0.029) (0.011) (0.012) (0.041)

P-values:

Jointly = 0 0.34 0.26 0.28 0.33 0.24 0.44

Passed=Failed 0.48 0.60 0.90 0.39 0.74 0.88

Observations 3,873 3,873 3,590 3,873 3,873 3,590

B. Non-Experimental Sample

National Board:

Achieved 0.000 0.000 -0.006 0.000 0.001 -0.010
 (0.001) (0.001) (0.010) (0.001) (0.001) (0.010)

Withdrew -0.005 -0.001 0.031 -0.004 0.000 0.024
 (0.002) (0.002) (0.019) (0.002) (0.002) (0.019)

Did Not Achieve 0.002 0.003 0.000 0.002 0.003 -0.001
 (0.002) (0.002) (0.018) (0.002) (0.002) (0.019)

P-values:

Jointly = 0 0.01 0.57 0.37 0.02 0.53 0.43

Passed=Failed 0.42 0.28 0.75 0.46 0.30 0.63

Observations 250,947 250,947 247,962 250,947 250,947 247,962

Note: All estimates control for school-by-year-by-grade fixed effects. Control variables include baseline math and reading scores (imputed to mean if missing) interacted with grade, dummies for missing scores interacted with grade, race/ethnicity (hispanic, white, black, other or missing), ever retained, title I, eligible for free lunch, homeless, migrant, gifted and talented, special education, english language development (level 1-5), and the means of these variables among all students in the class. Experimental estimates include only pairs of teachers who were randomized to classrooms, while non-experimental estimates include all other teachers teaching in school-grade-years with an NBPTS applicant. Standard errors (in parentheses) are clustered at the school-grade-year level.

Table 6. Impacts on Math and Language Arts Achievement

| Math Score | | | | Language Score | | | |
|-------------------|---------------|-------------|---------------|-------------------|---------------|-------------|---------------|
| End-of-Year Score | | Gain Score | | End-of-Year Score | | Gain Score | |
| No Controls | With Controls | No Controls | With Controls | No Controls | With Controls | No Controls | With Controls |

A. Experimental Sample

| | | | | | | | | |
|----------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| National Board: | | | | | | | | |
| Achieved | 0.070 (0.071) | 0.046 (0.049) | -0.010 (0.055) | -0.042 (0.049) | 0.084 (0.072) | 0.060 (0.043) | 0.014 (0.039) | -0.039 (0.045) |
| Withdrew | -0.036 (0.116) | 0.081 (0.078) | 0.115 (0.082) | 0.149 (0.069) | -0.092 (0.125) | 0.016 (0.073) | 0.147 (0.071) | 0.149 (0.081) |
| Did Not Achieve | -0.108 (0.097) | -0.173 (0.081) | -0.289 (0.096) | -0.355 (0.082) | -0.098 (0.104) | -0.134 (0.061) | -0.231 (0.104) | -0.210 (0.071) |
| P-values: | | | | | | | | |
| <i>Jointly = 0</i> | 0.510 | 0.060 | 0.020 | 0.000 | 0.430 | 0.050 | 0.030 | 0.000 |
| <i>Passed=Failed</i> | 0.140 | 0.010 | 0.010 | 0.000 | 0.150 | 0.010 | 0.030 | 0.050 |
| # Observations | 3,790 | 3,790 | 2,311 | 2,311 | 3,788 | 3,788 | 2,310 | 2,310 |

B. Non-Experimental Sample

| | | | | | | | | |
|----------------------|------------------|-------------------|-------------------|-------------------|------------------|-------------------|------------------|-------------------|
| National Board: | | | | | | | | |
| Achieved | 0.181 (0.034) | 0.009 (0.020) | 0.032 (0.021) | 0.007 (0.020) | 0.197 (0.037) | 0.006 (0.017) | 0.052 (0.019) | 0.003 (0.017) |
| Withdrew | 0.219 (0.083) | -0.056 (0.042) | -0.017 (0.047) | -0.049 (0.043) | 0.311 (0.092) | 0.012 (0.032) | 0.073 (0.042) | 0.014 (0.033) |
| Did Not Achieve | 0.070 (0.069) | -0.071 (0.040) | -0.049 (0.041) | -0.069 (0.040) | 0.130 (0.066) | -0.017 (0.026) | 0.023 (0.030) | -0.014 (0.026) |
| P-values: | | | | | | | | |
| <i>Jointly = 0</i> | 0.000 | 0.163 | 0.247 | 0.228 | 0.000 | 0.876 | 0.020 | 0.911 |
| <i>Passed=Failed</i> | 0.134 | 0.067 | 0.070 | 0.084 | 0.352 | 0.453 | 0.387 | 0.558 |
| # Observations | 249,213 | 249,213 | 249,213 | 249,213 | 249,499 | 249,499 | 249,499 | 249,499 |

Note: All estimates control for school-by-year-by-grade fixed effects. See notes to prior table for description of samples and variables included in specifications with controls.

Table 7. Estimates for Pooled Sample in Pre-Experiment Years 2000-02

| | Math Score | | | Language Score | | |
|--|-------------------------|-------------------|-----------------------|-------------------------|-------------------|-----------------------|
| | No Controls | With Controls | Student Fixed Effects | No Controls | With Controls | Student Fixed Effects |
| National Board: | | | | | | |
| Achieved | 0.145 (0.024) | 0.048 (0.014) | 0.039 (0.010) | 0.082 (0.023) | 0.004 (0.012) | 0.015 (0.009) |
| Withdrew | 0.040 (0.050) | -0.018 (0.030) | -0.001 (0.025) | 0.046 (0.051) | 0.008 (0.028) | 0.009 (0.020) |
| Not Achieved | 0.004 (0.040) | -0.051 (0.023) | -0.065 (0.017) | 0.033 (0.038) | -0.013 (0.020) | -0.032 (0.015) |
| Achieved* Experimental Sample | -0.043 (0.070) | -0.031 (0.036) | -0.006 (0.027) | 0.019 (0.075) | 0.027 (0.036) | 0.018 (0.027) |
| Withdrew* Experimental Sample | 0.098 (0.093) | 0.039 (0.054) | -0.005 (0.044) | 0.013 (0.098) | -0.038 (0.046) | -0.027 (0.034) |
| Not Achieved* Experimental Sample | 0.030 (0.093) | 0.073 (0.066) | -0.054 (0.044) | -0.088 (0.093) | -0.036 (0.053) | -0.086 (0.041) |
| Control* Experimental Sample | 0.088 (0.049) | 0.007 (0.030) | -0.009 (0.021) | 0.050 (0.046) | -0.030 (0.027) | -0.035 (0.019) |
| p-values: | | | | | | |
| <i>National Board Variables = 0</i> | 0.0000 | 0.0074 | 0.0000 | 0.0156 | 0.6993 | 0.0014 |
| <i>Exper Sample=Non Exp Sample</i> | 0.6567 | 0.4737 | 0.6610 | 0.8109 | 0.6072 | 0.1421 |
| <i>Exper Controls=Non Exp Controls</i> | 0.0726 | 0.8157 | 0.6572 | 0.2805 | 0.2618 | 0.0692 |
| <i>Passed=Failed</i> | 0.0072 | 0.0005 | 0.0000 | 0.1958 | 0.2991 | 0.0001 |
| Controls: | None | Student,Peer | Peer | None | Student,Peer | Peer |
| Fixed Effects: | School*Grade*Track*Year | | Student*School | School*Grade*Track*Year | | Student*School |
| # Observations | 405,563 | 405,563 | 467,282 | 402,523 | 402,523 | 463,211 |

Note: The outcome variables are the standardized Stanford 9 test scores used by the LAUSD in the 2000 through 2002 spring testing. Standard errors allow for clustering at the school by grade by calendar track by year level.

Table 8. Testing the Predictive Power of the NBPTS Sub-Scores

| | Math | Language Arts |
|--------------------------|--------------|---------------|
| <u>Hypothesis:</u> | | |
| All Subscores = 0 | <i>0.000</i> | <i>0.006</i> |
| Video Subscores=0 | <i>0.037</i> | <i>0.168</i> |
| Student Work Subscores=0 | <i>0.154</i> | <i>0.783</i> |
| DAE Subscores=0 | <i>0.032</i> | <i>0.140</i> |
| Assess. Ctr =0 | <i>0.042</i> | <i>0.170</i> |
| New NBPTS Scaling | <i>0.007</i> | <i>0.501</i> |
| Old NBPTS Scaling | <i>0.013</i> | <i>0.029</i> |
| Observations | 235340 | 235543 |
| R-squared | 0.63 | 0.7 |

Note: All of the above specifications were estimated with non-experimental sample in 2004 and 2005. NB applicants who had any missing subscores were dropped from the analysis. All specifications include full student and peer controls as well as school by grade by track by year fixed effects.

Table 9. Validating the NBPTS and Imputed Scaled Scores

| Math Score | | | Language Score | | |
|----------------------|---------------|--|----------------------|---------------|--|
| National Board Scale | Imputed Scale | | National Board Scale | Imputed Scale | |

A. Experimental Sample

| | | | | | | |
|--|-------------------|-------------------|-------------------|------------------|------------------|------------------|
| National Board: Applied | -0.032 (0.049) | -0.030 (0.046) | -0.017 (0.041) | 0.013 (0.043) | 0.005 (0.039) | 0.005 (0.034) |
| Standardized Scaled Score if applied (0 otherwise) | 0.113 (0.040) | 0.057 (0.044) | 0.069 (0.039) | 0.050 (0.034) | 0.000 (0.033) | 0.038 (0.029) |
| Teacher's Prior Standardized Value-Added in Subject | | 0.203 (0.039) | 0.200 (0.038) | | 0.245 (0.034) | 0.241 (0.035) |
| <i>Joint p-value on National Board Terms</i> | <i>0.02</i> | <i>0.43</i> | <i>0.20</i> | <i>0.22</i> | <i>0.99</i> | <i>0.40</i> |
| # Observations | 2858 | 2216 | 2189 | 2857 | 2215 | 2188 |

B. Non-Experimental Sample

| | | | | | | |
|--|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| National Board: Applied | -0.024 (0.020) | -0.031 (0.017) | -0.025 (0.017) | -0.009 (0.015) | -0.012 (0.015) | -0.003 (0.014) |
| Standardized Scaled Score if applied | 0.055 (0.017) | 0.007 (0.016) | 0.075 (0.016) | 0.033 (0.012) | 0.033 (0.012) | 0.050 (0.013) |
| Teacher's Prior Standardized Value-Added in Subject | | 0.186 (0.005) | 0.185 (0.005) | | 0.096 (0.004) | 0.096 (0.004) |
| <i>Joint p-value on National Board Terms</i> | <i>0.01</i> | <i>0.19</i> | <i>0.00</i> | <i>0.03</i> | <i>0.03</i> | <i>0.00</i> |
| # Observations | 247,818 | 225,668 | 225,552 | 248,107 | 225,942 | 225,826 |

Note: All estimates control for school-by-year-by-grade fixed effects, and include the control variables students and peers. Value-added was calculated for each teacher separately for math and language, based on observational data from 2000-2002. Samples are limited to teachers for whom prior value-added measures were available, and excluded those taught by teachers who applied for but then withdrew from National Board certification. The NB scaled score with "imputed weights" used data for the non-experimental sample to calculate the weights for summing up the sub-scores into a single scaled score. The NB scaled score, the new imputed NB scaled score and the prior value-added estimates have all been standardized and restated in standard deviation units.

Table 10. Between School Estimator: Are Comparison Teachers working with NBPTS Certified Teachers More Effective?

| Comparison Teachers | | National Board Applicants | |
|---------------------|---------------|---------------------------|---------------|
| Math | Language Arts | Math | Language Arts |

A. Experimental Sample

| | | | | |
|--|-------------------|-------------------|------------------|------------------|
| Standardized Scaled Score of National Board Teacher in same school-grade | -0.031 (0.043) | -0.038 (0.040) | 0.036 (0.028) | 0.033 (0.025) |
| # Observations | 1414 | 1412 | 1444 | 1445 |

B. Non-Experimental Sample

| | | | | |
|--|------------------|------------------|------------------|------------------|
| Standardized Scaled Score of National Board Teacher in same school-grade (averaged if more than one) | 0.010 (0.014) | 0.019 (0.012) | 0.085 (0.016) | 0.056 (0.011) |
| # Observations | 19537 | 19560 | 12911 | 12923 |

Note: All estimates control for year-by-grade fixed effects, and include the student-level control variables (but not peer-level controls). Samples are limited to teachers who teach in the same school and grade as a national board teacher, and excluded those taught by teachers who applied for but then withdrew from National Board certification. The NB scaled score is the score for the sample of non-applicants (the control sample) is for the NB applicant in their school and grade (averaged in the non-experimental sample when more than one applicant teaches in the same school and grade). The NB scaled score has been standardized and restated in standard deviation units.

Table 11. Bivariate Regression Coefficients from Pair-Level

| <u>Regressor:</u> | <u>Dependent Variable:</u> | | | Pre-Experimental Value-Added of NBPTS Applicant VA_{Pre}^{NB} |
|----------------------------|--|---|--|---|
| | Difference at Baseline $\bar{S}_{Pre}^{NB} - \bar{S}_{Pre}^{Comp}$ | Difference at End of Year $\bar{S}_{Post}^{NB} - \bar{S}_{Post}^{Comp}$ | Difference in Gain from Baseline $(\bar{S}_{Post}^{NB} - \bar{S}_{Pre}^{NB}) - (\bar{S}_{Post}^{Comp} - \bar{S}_{Pre}^{Comp})$ | |
| | (1) | (2) | (3) | (4) |
| $VA^{NB} - VA^{Comp}$ | 0.0270 (0.0561) | 0.1873 (0.0332)** | 0.1273 (0.0590)* | ----- |
| Scaled Score ^{NB} | -0.0004 (0.0011) | 0.0018 (0.0014) | 0.0031 (0.0014)* | 0.0014 (0.0007)* |

Note: Using the pairs of teachers for whom classrooms were randomly assigned, the above are from regressions using within-pair differences in student achievement, prior value-added and the scaled scores for NBPTS applicants. Each pair includes one teacher who applied for NB participation and a comparison teacher working in the same school, grade and calendar track. Because the dependent variable is the difference within each pair, there was one observation per pair. Heteroskedasticity-robust standard errors are reported.

Figure 1.

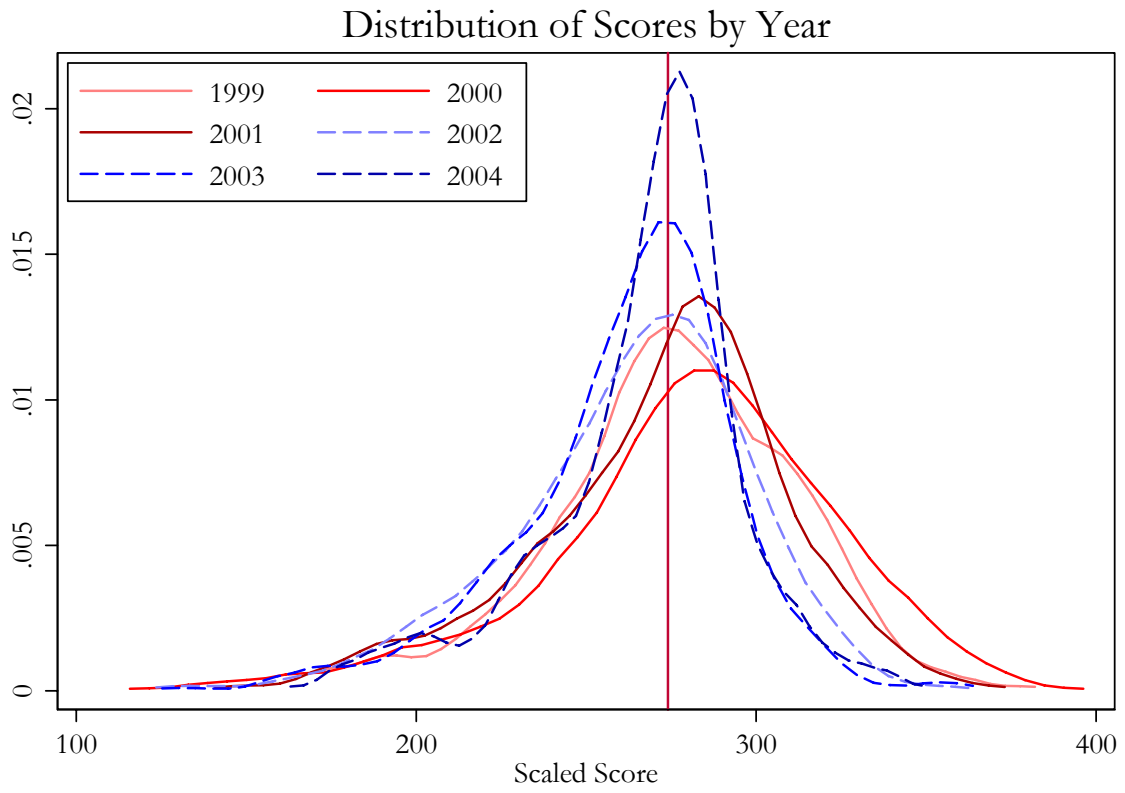


Figure 2.

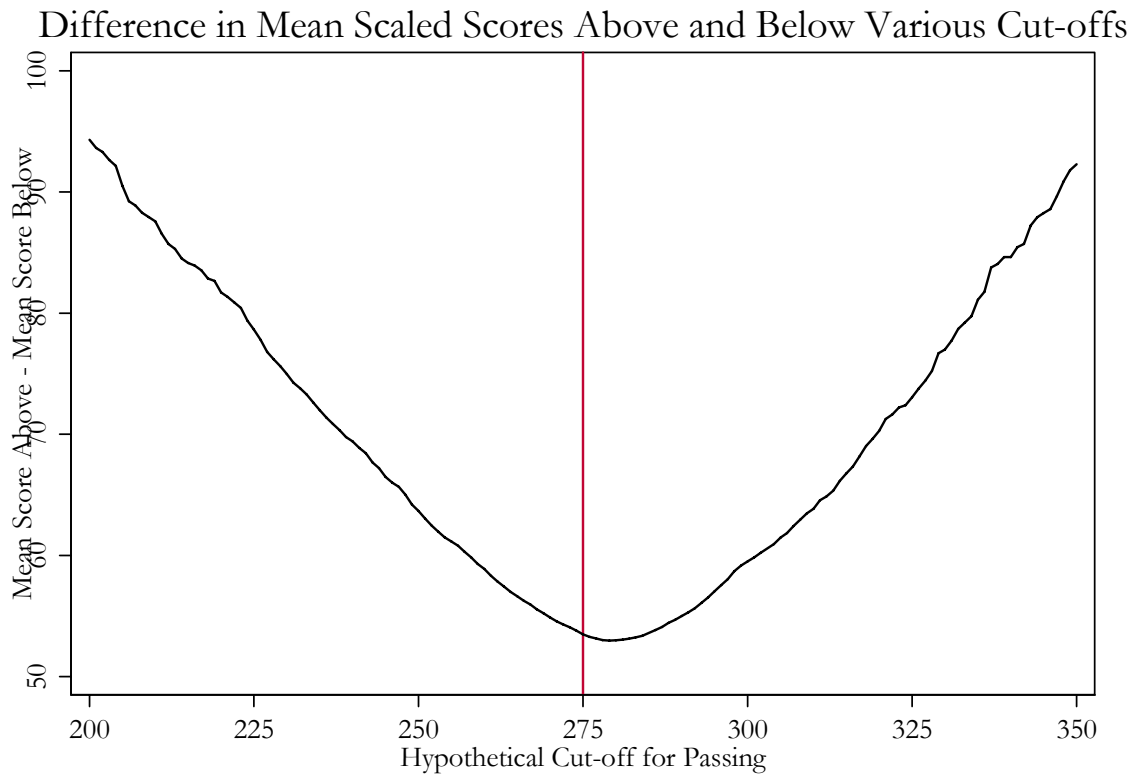
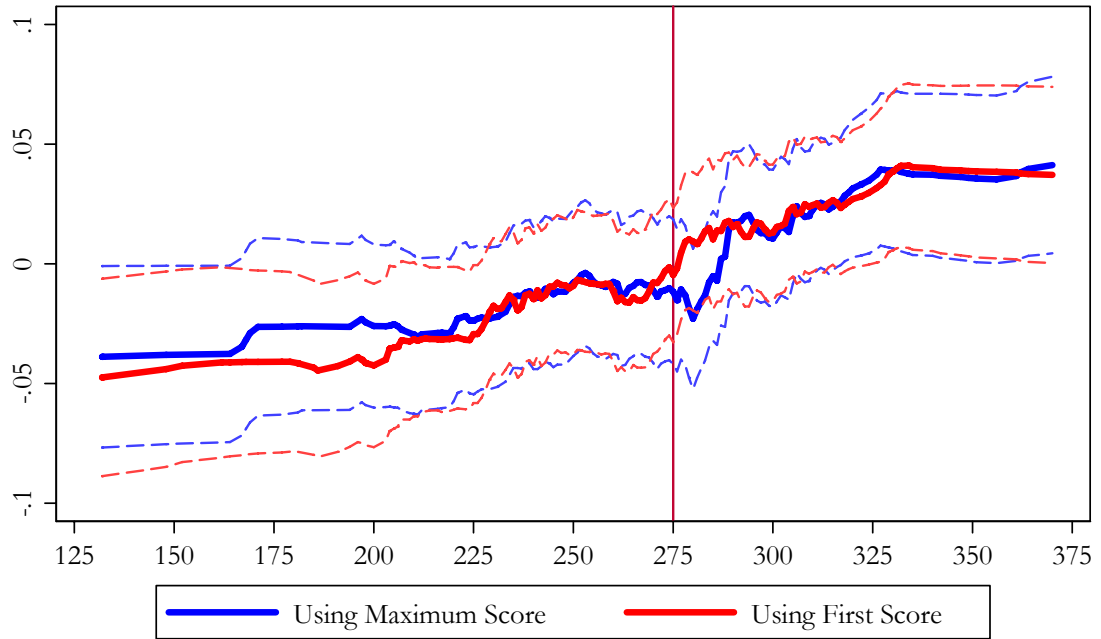


Figure 3.

Running Mean of Pre-Experimental Value-Added by NBPTS Scaled Score
First Scaled Score versus Maximum Scaled Score



Note: Running mean of 60 observations (30 to right and 30 to left). Passing score of 275 indicated.

Figure 4.

Within-Pair Differences at Follow-up and Pre-Experimental Value-Added
(One Observation Per Pair)

