

**Empirical Investigations of the Effects of National Board of Professional
Teacher Standards Certified Teachers on Student Outcomes**

A Report Prepared for the National Research Council

Daniel F. McCaffrey

Steven G. Rivkin

February 22, 2007

1. INTRODUCTION

Concern with the quality of education in the United States and beliefs that effective teachers were a key component to improving education led the Carnegie Forum on Education and the Economy, Task Force on Teaching as a Profession to recommend the creation of a board to “define what teachers should know and be able to do” and “support the creation of rigorous, valid assessments to see that certified teachers do meet those standards.” (Carnegie Forum on Education and the Economy, 1986). In response to that recommendation, a planning group for creating such a board was established and it eventually evolved into the board of directors for the National Board of Professional Teaching Standards (NBPTS, http://www.nbpts.org/about_us/background/history). Since its creation in the late 1980s, NBPTS has created standards for what accomplished teachers should know and be able to do and a criteria and procedures for a national voluntary system for certifying teachers who meet these standards. According to the NBPTS website over 50,000 teachers teaching over one million of the nation’s students are National Board Certified Teachers (NBCTs). Achieving this milestone of certification has been costly. Hundreds of millions of dollars were spent in developing the NBPTS standards and certification procedures. Millions more are spent each year by teachers applying for certification and by states and local school districts which provide financial rewards for NBCTs.

The NBPTS’s first policy statement *What Teachers Should Know and Be Able to Do* (http://www.nbpts.org/UserFiles/File/what_teachers.pdf) set forth the core competencies required of NBCTs. Criteria were then established for demonstrating these competencies for numerous teacher specialties as the basis for formal certification. The core competencies of accomplished teachers are given as five propositions in *What Teachers Should Know and Be Able to Do*. These propositions are:

1. Teachers are committed to students and their learning.
2. Teachers know the subjects they teach and how to teach those subjects to students.
3. Teachers are responsible for managing and monitoring student learning.
4. Teachers think systematically about their practice and learn from experience.
5. Teachers are members of learning communities.

The certification process is performance-based to measure a teacher’s teaching practice. The assessment includes an evaluation of four portfolio entries, three are classroom based including video recordings of teaching and examples of student work, and one demonstrates accomplishments outside the classroom. The certification assessment also includes an evaluation of the teachers’ content knowledge in their chosen certificate area. These evaluations include six 30 minute exercises administered at an assessment center designed for this purpose. Each teacher’s completed application is scored by a minimum of 12 teachers who have successfully completed intensive training and have been qualified for scoring based on their understanding of NBPTS standards and guidelines. To successfully complete the certification process, the candidate is required to earn a

minimum score on all of the sub-sections of the portfolio assessment and on various sub-tests taken at the assessment center.

Performance tests such as those chosen by the board are expensive to develop and to score. Thus, for teachers, the costs to take the examination are high, currently running about \$2,500.00.

Given the prominence of the NBPTS and the large amount of money spent on board certification and NBCTs, a growing number of studies have researched the relationship between certification and student outcomes. The National Research Council (NRC) was charged with evaluating the NBPTS certification process and reviewing the research on NBPTS certification and student outcomes. This paper contributes to that review. It has three purposes. First it reviews the eight studies from the existing literature on NBPTS certification and student outcomes. Second it reports some additional results that were requested from the authors in an effort to understand better the divergent findings in the existing work, and third, it draws some conclusions from both the literature and supplementary regressions on the relationship between teacher quality and NBPTS certification.

The next section develops a conceptual framework for the estimation of NBC effects that highlights the key identification issues. Section 3 summarizes the existing studies, focusing on differences in methods, samples, and findings. Section 4 reports the results of findings for Florida and North Carolina for a common set of specifications that we requested. An important question is whether the use of common specifications reduces the disparity in estimates of NBC effects in these respective states. Section 5 discusses the findings as a whole and offers preliminary conclusions regarding the effectiveness of the NBC program with respect to both its screening and career development functions.

2. CONCEPTUAL ISSUES

Analysis of the benefits of NBPTS certification requires both an understanding of the potential value of NBPTS certification, and the methodological issues that must be addressed in order to investigate empirically NBPTS certification effects. This section begins with a discussion of the potential value of NBPTS certification and then turns to a description of the key estimation issues.

2.1 Value of NBPTS Certification

Potentially NBPTS certification can provide a signal of instructional quality and participation in the process can lead to improvements in the quality of instruction. We begin with a discussion of the screening function of NBPTS certification in the context of the signaling model developed by Spence (1973). The key to NBPTS certification providing a signal of quality is that it is more costly for lower quality teachers to obtain certification, because the time it takes to compile a portfolio that would lead to certification varies inversely with teacher quality. The lower the correlation between cost and actual classroom effectiveness, either because the grading process is noisy or not well

aligned with performance in the classroom, the lower the average difference in effectiveness between NBCTs and other teachers. For example, if the NBPTS certification process was unable to distinguish portfolios demonstrating effective teaching from other portfolios then costs would not be greater for lower quality teachers.

If it is difficult to alter the signal by spending more time or money, the primary determinant of whether to apply is the probability of acquiring NBC conditional on actual quality. To the extent that teachers have information about their own effectiveness, lower-quality teachers would have lower expected rewards, because of a lower probability of being certified, and they would be less likely to apply. Again, low-correlation between performance in the classroom and expected rewards would result in more low-quality teachers applying for certification. As long as there is some uncertainty in the process or teachers information about their own efficacy, a higher reward will induce lower quality teachers to apply for NBPTS certification. Consequently the average difference between NBCTs and other teachers will depend in part upon the distribution of payoffs.

In addition to its screening value, the NBPTS certification process may have a direct effect on quality. Teachers typically spend hundreds of hours assembling a portfolio and participating in the program. On the one hand, this preparation and participation may lead to improvements in subject matter knowledge, classroom organization, or presentation. On the other hand, the time devoted to NBPTS certification may adversely affect the quality of instruction in the year of application for certification.

Given these two aspects of the certification process, there are at least three distinct research questions about NBPTS certification that these empirical studies might address.

These are:

1. Does the NBPTS screening process identify teachers who are more effective than other teachers?
2. Does the certification process change teacher effectiveness?
3. Given the NBC applicant pool and screening process, what is the average quality difference between NBC and non-NBC teachers?

The third question is clearly an important for policy makers. It allows administrators and other to determine if investing in certified teachers will tend to improve achievement and could support resource allocation. Moreover, knowing this value could allow for cost-effectiveness analysis of NBPTS certification and result in efficient policies. Assuming that the analysis of student test scores provides an accurate measure of teacher effectiveness, we can measure the average difference in effectiveness between teachers with certification and all others as the coefficient on an indicator variable for NBCTs.

The first and second questions examine the mechanisms that drive any average quality difference. Question 1) requires information on applicants who pass and those who fail in order to measure the effectiveness of board screening procedures. Question 2) requires information on applicants before, during and after the certification process in order to

identify any change in quality that accompanies certification. Longitudinal comparisons would provide the cleanest estimate but comparing the relative effectiveness of NBCTs before and after certification would also provide valid estimates under assumptions about the stability of the comparison teacher sample and the sample and the effects of NBCTs. This estimate would confound the learning effects of the actual application process with any effects that being certified has on teachers. For example, greater responsibilities for mentoring or teaching challenging students might be given to NBCTs and this could lower their effectiveness at increasing student achievement. Note that the estimate of the program induced change in effectiveness could be subtracted from the average gap between NBCTs and other teachers to estimate the portion of the NBCT non-NBCT differential that reflects pre-existing differences in teacher effectiveness.

The studies reviewed below use a variety of comparisons to understand the general value in NBPTS certification and to separate out the various components to that process.

2.2 Identification of NBPTS Certification effects

Most empirical analyses of NBPTS certification effects recognize the two mechanisms through which NBPTS certification is related to the quality of instruction but also the difficulty of separating the causal effects of NBPTS certification from other confounding factors. Studies thus differ both in the approach to parsing the various effects of NBPTS certification and the methods used to control for factors that might bias the estimates of NBPTS certification effects. This section provides a framework with which to evaluate the methods used to identify the causal effects of NBPTS certification.

An experiment that randomly assigned students to NBCTs and other teachers would provide an appealing approach to the identification of NBPTS certification effects, but to date such experimental findings are not available.¹ Consequently research must rely on observational studies in which the probability a student has a teacher with NBC depends on both the non-random sorting of teachers among schools and the non-random distribution of students among classrooms. The papers adopt a variety of methods to address these issues, and we specify additional empirical approaches that are reported and discussed in Section 4.

All authors recognize that the limited set of family background characteristics is unlikely to capture all confounding influences and consequently adopt a value added framework and in some cases include student, school, or both student and school fixed effects in the model. The term “student fixed effects” refers to including indicator variables in the model for each individual student in the data set, so that estimates depend on variations in scores (or gain scores) within students over time. Similarly, “school fixed effects” are

¹ Thomas Kane and colleagues have conducted a study that randomized students classes taught by teachers who had applied for NBPTS certification and those who had not. Although preliminary results of that study were presented at the NRC’s November 28, 2006 meeting, a report with those results is not available at this time.

implemented by including indicator variables for each individual school that equal one when a test score is from a student who was attending the school when tested. We now briefly discuss the purpose and potential inadequacy of various approaches used in the empirical work.

By controlling for prior achievement, the value added approach accounts for myriad family, community and school factors that affect achievement growth prior to the academic year in question. There are advantages and disadvantages to using gain as the dependent variable as opposed to including lagged score as a regressor with current test score as dependent variable (gain model imposes very strong assumptions about the rate of knowledge depreciation, while the lagged test score model is subject to biases due to measurement error). These differences must be kept in mind when interpreting the results.

Although school fixed effects account for systematic sorting of both students and teachers among schools by using only within school differences to identify the NBPTS certification effects, they do not address and may even exacerbate problems caused by the purposeful sorting of students and teachers within schools. For example, principals may match more difficult to educate students with NBCTs. If such student heterogeneity is not accounted for it may attenuate estimates of teacher quality differences or the effects of variables such as NBPTS certification that may in actuality be strongly related to quality.

Student fixed effects is an appealing approach to account for unobserved student heterogeneity. In this framework NBPTS certification effects are identified by differences in achievement gain when a student has an NBCT and when she has a non-NBCT. Thus the NBPTS certification effects are identified strictly by achievement gain differences for the same student.

The combination of student and school fixed effects has great appeal, but computational concerns limit its usefulness by requiring the spell approach described below. The problem with the approach is that only students who remain in the same school for at least two years contribute to the estimates, and this limits the numbers of students and teachers used to identify the estimates (notice that all teachers in the first grade offered in a school in the final year of the sample or in the last grade offered in a school in the first year of the sample are essentially thrown out).

In addition, there may be important time varying aspects of school or student quality that could bias the fixed effect estimates. For example, if the use of NBPTS certification is related to principal or even superintendent quality which changes across years due to turnover, school fixed effects may fail to account for important confounding factors. Similarly, if students are allocated to classrooms on the basis of annual performance, student fixed effects may fail to capture important time-varying student level factors. Finally, student effects on the classroom environment including any disruptive behavior may not be captured by student fixed effects or lagged achievement score. An example is a disruptive student with a high test score.

Recent work by Clotfelter (et al, forthcoming) uses a novel approach to address the problem of non-random classroom allocation within schools. Specifically, they identify schools in which observable student characteristics differ among classes no more than would be expected under random assignment. The drawback of this approach is that it has no power to detect schools that sort on criteria that is not observed by the econometrician.

One promising alternative to overcome the non-random sorting both within and between schools is to aggregate the data to the school by grade by year level and use the proportion of NBCTs as the measure of NBPTS certification (or the proportions of future, current, and previously certified teachers). Such aggregation eliminates biases introduced by purposeful classroom assignments within schools. Moreover, by reducing the size of the data matrix such aggregation enables the use of full sets of school by grade and school by year fixed effects. The remaining variation in proportion NBCTs is unlikely to be correlated with other determinants of achievement including unobserved student, community, or school factors (such as a new principal).

3. REVIEW OF EXISTING LITERATURE

At least ten different studies have used student test-score data to evaluate the effectiveness of NBPTS certified teachers. At the request of the NRC, we reviewed reports or papers documenting eight of these studies (Cavalluzzo, 2004; Vandervoot, Amrein-Beardsley, and Berliner, 2004; Goldhaber and Anthony, 2005; McCloskey et al., 2005; Sanders, Ashton, and Wright, 2005; Clotfelter, Ladd, and Vidgor, forthcoming; Clotfelter, Ladd, and Vigor, 2006; and Harris and Sass, 2006). The papers present conflicting views of NBCTs. Some papers (Cavalluzzo, 2004; and Vandervoot et al., 2004; Clotfelter, Ladd, and Vidgor, 2006) find significant and positive differences between certified and other teachers, some find no significant differences (Sanders et al., 2005), and others find that the results are sensitive to model specification, the comparison teachers and the timing of the comparison—before certification, after certification, or during the certification process—(Clotfelter et al., forthcoming; Goldhaber and Anthony, 2005; and Harris and Sass, 2006).

Two of these papers rely on small samples of teachers (Vandervoot, Amrein-Beardsley and Berliner, 2004; and McCloskey et al., 2005). Moreover both have methodological shortcomings which limit the value of their results. We provide a quick summary of these studies and two studies by Stone (2002, 2004), which also rely on small samples, before providing a more detailed review of the other papers.

3.1 Review of Studies with Small Samples

Vandervoot, Amrein-Beardsley and Berliner

The study by Vandervoot, Amrein-Beardsley and Berliner (2004) was funded in part by NBPTS and was the first author's doctoral dissertation. The quantitative study analyzed data from a sample of 35 self-selected NBCTs from 14 school districts in Arizona. The analysis compared the performance of these teachers' students on the Stanford 9, the state's accountability test at the time the data were collected, to the performance of other students in the same district. The sample included only students in grades 3 to 6. Differences between the students of NBCTs and other teachers were estimated using analysis of covariance (ANCOVA), where the outcome or dependent variable was the student's gain score (current year score less prior year scores) and the covariate was the student's prior year score.² The authors report separate results for reading, language arts and mathematics tests by grade and school year (1999-2000 to 2002-2003). They report the adjusted differences in group means converted to effect sizes by dividing each group adjusted mean by the within group standard deviation after controlling for prior scores.

Of the resulting 48 comparisons of NBCT students to other students from these districts, the authors report that ten were statistically significant and positive and the remainder were not statistically significant. There were no statistically significant negative adjusted differences and few negative differences overall.

There are, however, several methodological problems with this study. Foremost, the significance testing treated scores from students as statistically independent ignoring the potential for scores of students from the same classroom to be correlated. Failure to account for this clustering of students is likely to have resulted in standard errors of the estimated mean differences that are too small and for statistical significance to be overstated.

Another problem with the analysis is the authors' failure to account for the large number of comparisons being made and the increase risk of spurious findings that this created. Adjusting for a single prior year score is unlikely to have completely accounted for heterogeneity of students and to have removed the possible upward bias due to NBCTs students tending to be students more likely to score higher than other students regardless of their teacher. Finally, the authors do not account for the fact that the 35 participating NBCTs represent just 44 percent of the NBCTs in the 14 districts. It is possible that these teachers differ from other NBCTs in ways that affect their students' outcomes. There was no such selection of teachers who were not board certified; students from all of these teachers were included in the study. Consequently, the observed differences could represent differences between teachers who are willing to participate in the study and other teachers, rather board certification.

McColskey, Stronge, Ward, Tucker, Howard, Lewis, and Hindman

² Using gain scores with the prior year score as the covariate produces the same estimate of the difference between NBCTs and other teachers as an ANCOVA with the current score as the dependent variable and prior year scores as the covariate.

This study used data on NBCTs and other teachers and their students from two rural and one urban school districts in North Carolina. Because of limitations in the available test score data, the study restricted the sample to 4,632 fifth grade students' reading and mathematics scores on the state's accountability test. The resulting sample included 25 NBCTs and 282 other teachers.

For these teachers the study estimated teacher achievement indices (TAI) by adjusting student scores for prior performance, demographics and other student background variables and then averaging the adjusted scores for each teacher. The averages are multiplied by a factor less than one to reduce their sampling errors and these "shrunk" averages are used as the TAI. The study uses the difference between the TAI means for NBCTs and other teachers to examine differences in the outcomes for these two groups of teachers.

The differences are very small and not statistically significant. However, with such a small sample of NBCTs, the 95% confidence interval for the mean TAI for NBCTs is large and includes means that would be substantially different from the mean TAI for other teachers. Also, the use of the two stage procedure that adjusts student scores without accounting for their teachers NBCT status could result in over adjusting student scores and possibly biasing differences between the two groups of teachers toward zero.

Stone

Stone (2002) studies 16 NBCTs in Tennessee. At the time of this study, a total of 40 NBCTs taught in the state but only 16 had taught students in grade 4 to 8 and had sufficient data to receive Tennessee Value-Added Assessment Scores (TVAAS) teacher performance scores. The study compared TVAAS teacher performance scores to standard values of 115 (for exemplary teaching) and 85 (for deficient teaching), which indicate that the teacher added 115 or 85 percent of the average growth to his or her students' performance on the states accountability test. Stone reports that:

Considering the 16 teachers collectively, there are 123 teacher-by-subject-by-year teacher-effect scores. Only 18 (15 percent) of these scores reach the "exemplary" or "A" level and 13 (11 percent) would be designated as "deficient" and given a grade of "F."

Clearly the NBCTs did not score uniformly high on this metric. However, with so few teachers in the sample, it is difficult to make any conclusions about national board certification from this study, other than to conclude that the study does not suggest that NBCTs are uniformly accomplished on TVAAS teacher performance metric.

We now turn to the remaining studies which rely on larger samples of teachers and more sophisticated analytic approaches to account for possible differences between the propensity to learn of students taught by NBCTs and other teachers. We first provide detailed summaries of each paper and then compare the findings from the various papers.

3.2 Review of Studies with Large Samples

Cavalluzzo

The study by Cavalluzzo (2004) was one of the first large scale investigations of differences between the outcomes of students taught by NBCTs and other students. The study used data on ninth and tenth grade mathematics students from the Miami-Dade County schools for the 2000-2001 to 2002-2003 school years. The sample included 107,997 students and 2,137 teacher years. It includes all the NBCTs teaching the selected grades during the chosen school years. Student scores on the state's end of grade accountability test (the Florida Comprehensive Assessment Test, Sunshine Standards Tests, FCAT-SSS) provide the measure of student outcomes.

The paper used education production function methods to study difference between the outcomes of students taught by NBCTs, current NBPTS applicants, teachers who applied for but did not receive NBPTS certification, and other teachers. The covariates used in the models are detailed student background variables including grade-level, age, gender, race-ethnicity, English language proficiency, participation in free or reduced price meals, grade retention, gifted status, special education status, school suspensions, days absent, GPA, math effort and conduct, and whether the students math class was above, below or on grade-level. The models also included the teacher background variables: whether or not the teacher is teaching within subject area of certification, the teacher's salary step (as a measure of years of experience), the teacher's certification status, whether or not the teacher has a graduate degree, and the selectivity of the teacher's undergraduate college or university. The production function used a linear model that included these variables and the student's prior year math score to predict current scores. Some models also include variables measuring school attributes and others included school fixed effects, i.e. indicator or dummy variables for each school that equal one if the student attended the school and zero otherwise. Scores from 9th and 10th grade students were combined into a single data set and fit to a single model.

The study finds that after adjusting for all the variables mentioned above, the test scores of students whose teachers were NBCTs were statistically significantly higher than students whose teachers had no participation with NBPTS. Similarly, students whose teachers were currently NBCT applicants scored statistically significantly higher than students whose teachers had no participation with NBPTS; whereas, students whose teachers applied for NBPTS certification but failed to be certified scored statistically significantly lower than students whose teachers had no participation with NBPTS. For the author's preferred model, the effect sizes were about 0.07 (standard deviations) for certified teachers, 0.02 and -0.02 for currently applicants and teachers who failed to receive certification. These results were relatively insensitive to variations in the model including the use of student fixed effects instead of using prior year scores as a covariate.

The large sample of students and thoroughness of the models gives the results of this study more credibility than the previously reviewed studies, but the analysis does have some limitations. First the analysis does not account for the fact that student test scores

are nested within classes, within schools. Often scores for students from the same classroom or school are positively correlated and ignoring the possible correlation could result in standard errors that are too small and significance levels that are too low resulting in spurious statistical significance. Given that teacher status is a teacher level variable, ignoring the nesting of students within a teacher could be particularly problematic. In particular the sample contained over 100,000 students but only 61 teacher years for NBCTs, only 101 teacher years for teachers whose certification was pending, and only 18 teacher years for teachers who applied but failed to receive certification or withdrew their applications. Even if the correlation among scores from students within the same class is small the effective sample size for testing effects could be substantially smaller than that assumed in the analysis. Given that the effect sizes are small, it is very likely that many would not be statistically significant.

A second concern is that using prior year scores as a covariate can result in bias when the residual errors in test scores are serially correlated. The analysis with student fixed effects, which would not be susceptible to this bias also find results that match those of models that use prior score as a covariate. It is not surprising that the point estimates are lower in the student fixed effects models, because they impose the unrealistic assumption that knowledge fully depreciates each year such that past experiences have no effect on current achievement. Nonetheless, the effects are likely not to be significant with correct standard errors. All in all, the estimates suggest positive effects, but the evidence is weak.

Another problem is that the model does not account for the course contents and NBCTs might not be teaching courses with the same content as other teachers. This could result in confounding of content and NBCTs effects. There is no discussion of course content so the possible extent of bias cannot be assessed. Course content is particularly problematic with high school mathematics students because the content is highly differentiated across courses and the tests are not course specific.

Sanders, Ashton, and Wright

Sanders, Ashton, and Wright used data for a sample of third to eighth grade students from Charlotte-Mcklenberg and Wake County School Districts to study the differences between NBCTs and other teachers. The study included both mathematics and reading scores from the 1999-2000 to the 2002-2003 schools years resulting in over 130,000 student test score used for each subject. The study divided teachers into five groups, NBCTs, future NBPTS candidates, NBCT applicants who failed to be certified, and teacher with no NBPTS involvement. For mathematics there were 281 teacher years for certified teachers, 117 for failed applicants and 277 for future applicants. For reading the numbers were 206, 177, 253 teacher years for NBCTs, failed applicants and future applicants respectively.

The authors fit four models for each subject. In two models they used current year score on the state's end-of-year test as the outcome or dependent variable, and in the other two models they used the gain score (prior year score less the previous year score) as the outcome variable. For both subjects (mathematics and reading) and each outcome (level score or gain score) one model included random effects for teachers and the other they

did not. Including teacher random effects accounted for the nesting of students within classes and should provide substantially more accurate standard errors than the models that ignore this nesting. Models with the current score as the outcome included prior year mathematics and reading scores as covariates. All models also controlled for students' gender and race-ethnicity, teachers' years of experience and indicator variables three levels of participation with NBPTS (NBCT, failed applicant, and future applicant). The study fit models separately by grade.

The study found significant difference between the student outcomes for NBCTs and teachers not involved with NBPTS for grades 5, 6, 7 and 8 in both mathematics and reading for at least one outcome specification. For mathematics none of the differences were significant in models that include random teacher effects. In reading differences were significant in models that do and those that do not include teacher random effects except in grade eight where the effects were only significant in models that included teacher random effects.³

The paper does not report comparisons of the other groups to teachers with no NBPTS experience.

Given that the samples of teachers were relatively small, especially in grades 6, 7, and 8, the major contribution of this paper is the illustration of the sensitivity of results to the calculation of standard errors. Appropriate treatment for the nesting of students in classrooms generally resulted in substantially larger standard errors, which suggests that analyses that do not account for such nesting produce downward biased standard errors which raise the probability of reporting significant NBC effects in error.

One of the possible shortcomings of this study is that it gives up power to detect differences by analyzing the grades separately. This is particularly problematic given the small numbers of teachers in grade 6, 7, and 8. However, as shown in Figure 1, the differences between NBCTs and teachers not involved with NBPTS vary considerably across grades with some of the largest differences between grades 4 and 5, which have the largest samples. Thus, even if the data were pooled across grade it is unlikely that a strong and significant difference would exist for NBCTs.

³ It may be somewhat surprising that the only significant effects in eighth grade math occurred in models that included teacher random effects and including random effects tends to increase standard errors. However, it also can re-weight the contribution of individual teachers to estimated effects and change point estimates, especially with small sample sizes as in the eighth grade sample of teachers.

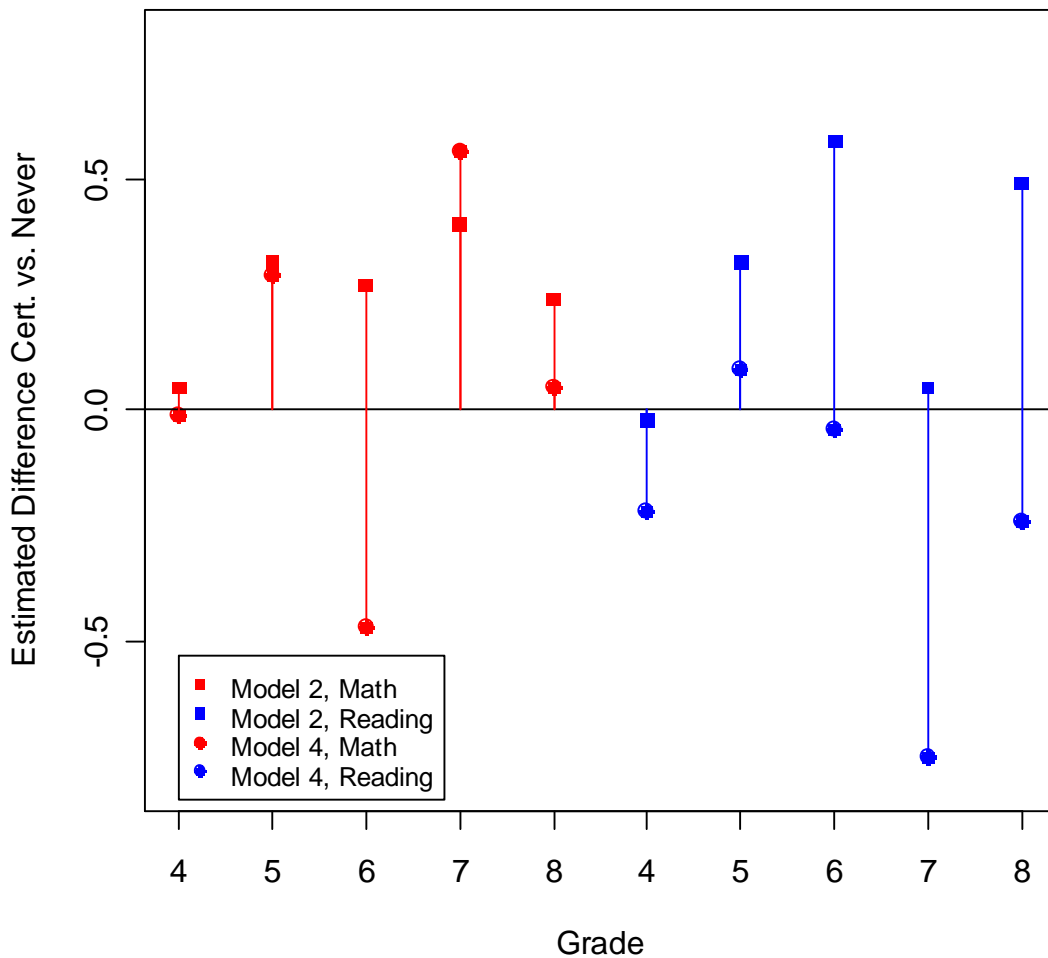


Figure 1. Estimated differences between NBCTs and teachers no involved with NBPTS by grade, subject, and model from Sanders, Ashton, and Wright (2005). Model 2 uses student achievement as the outcome and includes teacher random effects. Model 1 uses gain score as the outcome and includes teacher random effects.

Goldhaber and Anthony

Goldhaber and Anthony used mathematics and reading data for students in grades 3 to 5 from the entire state of North Carolina for the 1996-1997 to the 1998-1999 school years. For grade 3, the study included scores from both fall and spring testing, whereas in grades 4 and 5 it included only data from spring testing. All test score data were from the state's accountability tests. Overall the dataset included 611,517 student observations with both pre and post test scores. For grades 4 and 5 pre-test scores are from the prior year of testing; for grade 3 the pre-test is the fall test.

The study used education production function methods for student gain scores to estimate the differences between teachers with differing involvement with NBPTS. The production functions included the detailed student level variables: grade-level, race-ethnicity, gender, free or reduced price meal program status, English language proficiency, and special education status. The models also included the teacher variables: age, race-ethnicity, years of experience, advanced degrees status, credential status, score on a standardized test of teaching converted to z-scores so that scores from different tests can be treated as a single variable; school level variables: total enrollment, percent minority, student-teacher ratio, percent free or reduced price lunch students, and district level variables: total enrollment, per-pupil expenditure, percent of expenditures spent on instruction, urbanicity, starting salary for teachers, median housing value, percent in the community with at least a B.A. degree and an indicator variable for year of test.

The authors considered four specifications for their models. The first included all the covariates describe above except the teachers standardize test score. The second included all the covariates, the third replaced all the school and district variables with school fixed effects and the fourth replaced all the student covariates with student fixed effects. The model with student fixed effects did not include school fixed effects. The authors included three variations of Model 2 (the model with all the covariates and no fixed effects). The first included an indicator variable for whether or not a teacher was currently an NBCT and a separate indicator variable for whether or not the teacher would be an NBCT in the future, but was not currently an NBCT. The second variation to Model 2 separated current NBCTs into those in their first year of certification and those certified more than a year ago. This variation also separated future NBCTs into current applicants and other future NBCTs. The third variation to Model 2, allowed for the study of the application process by including variables for future applicants, current applicants and past applicants. Past applicants included current NBCTs and NBCT applicants who were not certified. Separate models were fit for reading and mathematics.

The authors found that compared to teachers who never applied to NBPTS or applied and did not receive certification, NBCTs' students made significantly higher gains in reading but not in mathematics. They also found that the students of teachers who would be certified in the future made consistently higher gains than the students of other teachers. This result held for both reading and mathematics and regardless of the model's specification of the comparison group of teachers. In addition, the study found that current applicants' students made lower gains on average than students of other teachers and this result held for all the mathematics models and nearly all reading models. Thus, there is consistent evidence that the teachers certified by the NBPTS were teachers who prior to application for certification taught students making larger gains in achievement than comparable students and that during the application process, applicants' students did not perform as well as other students.

However, the story for currently certified NBCTs was less clear. Students of NBCTs in their first year of certification had larger gains than similar students in other teachers' classes. This result held for all mathematics models and all reading models except for the model with student fixed effects. However, the results for teachers who had been

certified for more than a year were very inconsistent. In general the differences were positive for reading but not significant and they were negative for mathematics but significant only in the model with student fixed effects.

In general, the results were invariant to the inclusion of teacher test scores as a covariate and relatively insensitive to the inclusion of school fixed effects as opposed to school level covariates. The models were much more sensitive to the inclusion of student fixed effects and this could have been due to changes in the sample size.

The study's large sample of students and its detailed data for removing potential confounds enhance the importance of this study's results; however, there are some important concerns with this study. First, the authors do not account for the clustering of students in classes. As demonstrated by Sanders, Ashton and Wright (2005), controlling for such clustering can substantially increase standard errors and the tests that fail to control for clustering could be biased in favor of finding effects. Thus, we need to be cautious in interpreting statistical significance in the current study. The concern about bias in the statistical tests is exacerbated because many of the effects for NBPTS applicants and NBCTs are very small and the models generally account for very little of the variance in gains except for the models with student fixed effects.

Another concern with this paper stems from the results on differences of the effects NBCTs at different times in the application process. Ideally such results would use longitudinal data on teacher to determine differences in performance during the application process so that differences could be attributed to the process not the teacher sample. However, Goldhaber and Anthony (2005) were unable to use longitudinal data on teachers for their analyses. Thus, the sample of teachers who were certified for just one year did not contain the same teachers as the sample of teachers who had been certified more than one year and the sample of teachers prior to certification might not have included all teachers in the certified group. Differences in student outcomes among the groups of teachers at different stages of the certification process could have resulted from sampling error among the teachers. Moreover, differences in the groups were not tested formally and it is difficult to determine exactly how many individual teachers are in each group.

A final limitation is that the paper never compares NBCTs to applicants who failed so it is not possible to evaluate the efficacy of the screening process.

Clotfelter, Ladd, and Vigdor (forthcoming)

This study again uses mathematics and reading data for the entire state of North Carolina but focuses only on grade 5 student outcomes from the 1999-2000 school year, because the authors had test score data only for grades four to eight and teacher student links were questionable in secondary grades. The study used the state's accountability test as the outcome variable for the roughly 68,000 students in the data set.

The authors used a series of linear regression models with the 1999-2000 score as the dependent or outcome variable and different set of student variables to control for for the non-random assignment of students to classroom. However, every model included the following teacher-level variables: race-ethnicity, gender, years experience, quality of the undergraduate institution, an indicator for having an advanced degree, score on teacher licensure test, class size and an indicator for NBCTs.

This paper took a novel approach to control for the nonrandom assignment of students to teachers. First rather than use gain scores as the outcome, the authors use the prior year score as another covariate. However, they note that if test scores are serially correlated then this adjustment could results in biased estimation and they argue the direction of the bias cannot be determined. If other covariates remove the serial correlation, then the coefficients for teacher variables should be invariant to the inclusion of prior year test scores in the model. Thus, the authors fit a series of models that started with no controls for students, added basic student demographic variables (gender, race, free or reduced price lunch status), and then added extended student variables (computer use, time spent free reading, time spent watching TV, parental education, and time spent on homework). Finally they added school fixed effects to the model. At each stage, they fit models with and without prior year test scores. They stopped expanding the model after adding school fixed effects because the coefficients on teacher variables were generally invariant to adding prior year test scores to this model.

The authors explored another novel method to account for nonrandom assignment of teachers within schools. For every school in the state they determine if the variation between classrooms of six student-level variables was greater than expected under random classroom assignments. The variables used in this test were: the three student level demographic variables listed above, prior year tests categorized to above or below the state average, an indicator for whether or not the student attended the school in the prior year, and parental education as reported in the prior year. To increase power to detect nonrandom assignment, they pooled data from students in grades three to five when possible. Having completed these tests, the authors failed to reject the null hypothesis of nonrandom assignments, in 521 of the states 1,160 schools educating fifth graders. They then used only these schools and refit their final model (i.e., the model with the complete set of student variables and school fixed effects both with and without prior year scores) to both the mathematics and reading data to estimate the effects of teacher qualifications on student achievement.

The analyses in this paper accounted for the clustering of students in classrooms by using empirical sandwich standard error estimates (Liang and Zeger, 1986) with classroom as the nesting factor.

The authors found that in their final model fit to data from all schools, students taught by current NBCTs scored higher than students taught by all other teachers on reading achievement but not on mathematics. The effect was very small and it was not significant when the authors restricted the sample to those schools that appear to use roughly random class assignments. In that model, the coefficient for NBCTs is six times

smaller than it was in the model fit to data from all the schools. The loss in significance is not just from a loss of power resulting from a smaller sample. Hence this model provides little evidence for better student achievement for NBCTs students when compared to similar students taught by other teachers.

The results of this paper are compelling due to the novel methods used for controlling for nonrandom assignment and the large sample sizes. However, it does have some shortcomings that need to be considered when interpreting its findings. First, there are a large number of teacher variables included in the models. Thus, the coefficient for NBCTs is estimating the difference between an NBCT and other teachers when the teachers have similar characteristics including quality of undergraduate training, years of experience, advanced degree status, and licensure test score. These other attributes might be the source of differences between NBCTs other teachers. If this were the case, the NBPTS certification process might provide valid indicator of better teachers but it might not contain information that not could be obtained from other sources (years of experience for instance). On the other hand, it seems unlikely that all experienced teachers meet the criteria of the NBPTS, so it seems reasonable that certification if meaningful should find significant differences among teachers with the same experience.

Although the restriction to schools with apparent random classroom assignment on observed student variables is novel and is likely to help reduce the chance of potential bias, its value might very limited. The models account for these student-level variables so they are unlikely to be a large source of bias even among schools that use purposive assignments. Moreover, even if the schools randomly assign students on the basis of these variables they might purposively assign student on other characteristics such as classroom behavior that are not observed.

Clotfelter, Ladd, and Vigor (2006)

In this paper the authors model the relationship between a variety of teacher characteristics and student achievement test scores using data on all North Carolina students in grades 3, 4, and 5 for the 1994-1995 to the 2003-2004 school years. The number of students with scores increases from about 88,000 students per grade in 1995 to about 102,000 students per grade in 2004. The source of the growth in the sample is not discussed in the paper. The analyses used the state's end of grade test scores for students in each grade. For each grade and year scores are standardized to have mean zero and standard deviation one. Analyses that required prior year score were restricted to grades 4 and 5.

The study again used a production function approach and a series of alternative specifications for the model. To motivate their model, the author first introduce a simplified model for student achievement with the strong assumptions that the effects of teachers quality on student achievement was the same at every grade level and was the constant across all years of the study. In addition, they assumed that these effects decay at a constant rate every year. This yields a structural model for current year test scores as an additive linear function of the prior achievement score and current year teacher inputs.

The authors used this model to motivate five more complex models that they then fit to the data to estimate the effects of various teacher attributes on student achievement.

The first model was a simple value-added model with current year score as the outcome or the dependent variable, and the explanatory variables in the model included prior year score, time-invariant teacher characteristics (gender, race, competitiveness of the teacher's undergraduate institution, and teachers scores on a licensure test), time-varying teacher characteristics (years of experience, advanced degree status, licensure, and NBPTS certification status), classroom characteristics (class size, percent non-white students, percent of student receiving subsidized lunch, average parental education of students, average prior year scores), time-invariant student characteristics (race, gender, and age in grade 3), time-varying student variables treated as time-invariant because of data limitations (reduced price meal status and level of parental education), and time-varying student variables (repeating a grade, transfer into a school, and transfer as part of school system structural change, e.g., movement from elementary to middle school).

The authors extended this model by adding school fixed effects, so that the effects of teacher characteristics were measured by variation within schools and differences in the student populations across schools were not confounded with the estimates of the effects of teacher characteristics. The third model used student gain scores (current year score less prior year score) as the dependent variable rather than using level score as the dependent variable. This model did not include prior year score as a covariate. The fourth model returned to using current year achievement level as the dependent variable but replaced student prior year test score and student time-invariant variables with student fixed effects. The fifth model used student fixed effect with gain scores. The authors discussed the potential bias with each model and presented coefficients for the teacher variables from every model. All models were fit separately for mathematics and reading using all available student data. Consequently, Model 4, which does not use lagged scores, includes third grade scores as outcomes but no other model does.

The primary model specification included an indicator variable for whether or not a teacher is currently a NBCT. For mathematics, the coefficient is statistically significantly positive for every model. The coefficients range from 0.018 to 0.028 but most estimates are very close to 0.02. Given that scores were scaled to have standard deviation 1.0, these coefficients imply that students in NBCTs' classes scored about two percentage points of standard deviation unit higher than similar students in other classes. This is an effect size of about 0.02. Compared to other effects from the models, the effect of NBCTs on mathematics scores was about one third the size of the effect of being a new teacher or one fourth the size of the effect of student race and one tenth of the effect of having a parent who was a high school dropout. However, the effect for NBCT was comparable to or larger than nearly all the other teacher quality measures except initial experience.

The effect for NBCTs on reading was again positive with coefficients ranging from 0.012 to 0.018 with an average across models of about 0.014. The coefficients were

statistically significant in all models, except for Model 5, which used gain scores as the dependent variable and included student fixed effects.

Using the models with student fixed effects, Models 4 and 5, the authors conducted a more detailed evaluation of NBPTS certification. Following Goldhaber and Anthony (2005), they explored the signaling effects of NBPTS certification by estimating the effects student outcomes of being in a classroom taught by a teacher who was not currently an NBCT but who would be certified during the span of the study. They considered teachers two years prior to certification, one year prior to certification, the year of certification, and one or more years after initial certification. The authors found that for mathematics using Model 4, that the effects were largest two years prior to certification and post certification with a dip in effects the year before certification (the application year) and the first year of certification. However, with Model 5 the effects were largest for teachers prior to certification and smallest in the two years post certification. Moreover, all of the effects for Model 5 in this secondary analysis were two or more times larger than effect of NBCTs in the model that included a single indicator for current NBCTs. The size of the effects in this secondary model is hard to interpret given that effects for teachers post certification should be similar to the single effect for NBCTs.

For reading, Model 4 suggests that for teachers who are certified sometime during the study, their students scored highest relative to other students when the teachers were two years prior to certification. The effects get smaller with every year of certification staging, so that the effects for NBCTs post-certification were less than half as large as the effects two years prior to certification. This pattern did not repeat with Model 5. In fact, with Model 5, students of certified teachers did best when the teacher was two years prior to certification and during the year of certification. Thus, for both reading and mathematics the results of this secondary analysis were highly sensitive to model specification and inconsistent with the simpler model formulation that included a single indicator for current NBCTs. These analyses thus yield unstable estimates that need further investigation.

Although each model specification can potentially yield biased results, the consistency of effects for NBCTs across multiple models for both mathematics and reading provides compelling evidence that the cohort of NBCTs in North Carolina between 1995 and 2004 raised achievement test scores more than other teachers.

One shortcoming of the paper is the fact that the authors do not use the longitudinal data on the teachers to study how the same teacher's students score as the teacher's NBPTS status changes. This could provide more interpretable measures of NBCT effects than the comparisons that compare teachers prior to certification to other teachers.

Harris and Sass

This study used data from the entire state of Florida from the 1999-2000 to the 2003-2004 school years to model growth in student achievement and test for differences

between NBCTs and other teachers. The study used data from students in grades 3 to 10 on both mathematics and reading for both the state's Florida Comprehensive Assessment Test (FCAT) norm referenced test (FCAT-NRT), and the state's criterion referenced FCAT Sunshine State Standards Test (FCAT-SSS). The sample included over 2.2 million test scores for each of mathematics and reading on each test.

This study again used a production function approach. The study used gain scores in achievement (current year score less prior year score) as the outcome or dependent variable. The model included as independent variables: school fixed effects, student fixed effects, time-varying students variables (number of schools attended in the year, an indicator for making a structural move (e.g., switch from elementary to middle school or middle to high school), and an indicator for making a non-structural move), peer variables (the classroom proportion female, black, or undergoing a structural move, the classroom mean age, and class size), and indicators for teachers NBCT status. The authors report that, to avoid computational problems, they included student by school or "spell effects" for each period a student is in a different school rather than including separate fixed effects for each school and separate fixed effects for each student.

The authors fit two primary models. The first included a single indicator variable for whether or not a teacher was ever NBPTS certified during the span of the data. The second model estimated separate effects for teachers ever NBPTS certified during three periods of the NBPTS process: the years prior to application, the year of application, and the years following certification. The model included separate indicator variables for teachers in each group. Both models were fit separately to reading and mathematics gain scores and for both the FCAT-NRT and FCAT-SSS tests.

The study finds no significant difference between student achievement gains when they were taught by a teacher who was or would become a NBCT. This result held for mathematics and reading for both the FCAT-NRT and FCAT-SSS. However, the two tests (FCAT-NRT and FCAT-SSS) yielded somewhat different stories about NBCTs. The estimates for the FCAT-NRT were negative for both reading and mathematics whereas the estimates for FCAT-SSS were positive for both subjects.

Disaggregating NBCT effects by period yields a more complex set of results. Gains on the FCAT-NRT mathematics test were statistically significantly greater for students whose teachers would someday be NBPTS certification. However the gains on these tests were negative for students whose teachers were currently applicants for NBPTS certification (and would be awarded certification) and for students whose teachers were currently NBCTs. For mathematics FCAT-SSS, students in each group made greater gains than students whose teachers would never be board certified in the span of the data, but none of the differences were significant. In reading, students whose teachers were not currently NBCTs but would someday be awarded NBPTS certification and students whose teachers were currently NBCTS made greater gains on the FCAT-NRT than students whose teachers would never be NBPTS certified during the span of the study; however, neither difference was statistically significant. The gains on FCAT-NRT reading tests are negative, but not significant, for students whose teachers were currently

applicants for NBPTS certification (and would be awarded certification). The FCAT-SSS reading gains are significantly higher for students whose teachers would someday apply for and be awarded NBPTS certification and for students whose teachers were currently NBCTS than for students whose teachers would not be awarded certification during the span of the study. The FCAT-SSS reading gains were negative but not significant during the application year of future NBCTS.

The authors also considered many alternative models restricted to teacher from limited grades or years of certification and the story became even less consistent and more difficult to interpret. In general, this study provides no convincing evidence about NBCTS

As with the studies by Clotfelter, Ladd, and Vigor (2006, forthcoming) and Goldhaber and Anthony (2004), the large sample size and aggressive modeling to reduce the potential for bias from nonrandom assignment of students, give this paper more credibility than other studies but there remain some important concerns with this study. First, the paper makes no adjustment for the clustering of students within teacher and as discussed before this could result in biased significance tests which overstate the significance of results. Second, as discussed above, the complex adjustments of including student by school or spell fixed effects results in many students and teachers being excluded from the analysis. For example, students must be in a school for a minimum of two years to contribute to estimation of teacher effects. Similarly some teachers will be excluded from the analysis; for instance, teachers who teach only during the last year and teach third, sixth or tenth grade will be excluded. Moreover, these adjustments yield consistent estimates only under many assumptions including the assumption that students' achievement scores are growing at student-specific rates. Thus, it is possible that restrictions to the sample and estimation error that results from using many fixed effects lead to the inconsistency of the study's results.

3.3 Summary

Across these many studies and the variations of models explored within individual studies, the findings on NBCTS are mixed. Several studies find significant positive effects when comparing students of NBCTS to other teachers (Cavalluzzo, 2004, Goldhaber and Anthony, 2005, Sanders, et al, 2005, Harris and Sass, 2006, and Clotfelter et al, 2006, forthcoming). Other studies find negative effects when comparing students of NBCTS or NBPTS applicants to other students (Cavalluzzo, 2004, Goldhaber and Anthony, 2005, Sanders, et al, 2005, Harris and Sass, 2006, and Clotfelter et al, forthcoming).

There are several factors that might contribute to the heterogeneity in estimated effects. First the studies compare NBCTS and NBPTS applicants to different comparison groups of teachers and these differences in the comparison groups could be contributing to inconsistencies in the findings about board certification. Second across studies and even within a single study multiple models are used to estimate the effects of board certification. Different models inevitably produce different estimates both because of finite sampling error and different controls for confounding factors. Finally, the estimates

are made with data from different cohorts of teachers in different states, teaching different grades and in different years. There could be cohort effects in the teachers and there could be various context effects that could influence the performance of NBCTs relative to other teachers. We will now consider more specifically how estimates vary with each of these factors.

Differences in NBPTS certification parameterization

As discussed in Section 2.1, there are multiple questions about NBPTS certification that can be explored empirically by comparing different groups of teachers. There are general comparisons of NBCTs to other teachers and comparisons to distinguish the signaling effect of certification from the effect of the certification process and the effects of being certified. Following Goldhaber and Anthony (2005) most authors attempted to distinguish the signaling component of NBPTS certification from the effects of the process through a series of comparisons of different groups of teachers. To facilitate comparisons among studies we need to understand the contrasts different studies are making. We begin by describing the stratification of teachers considered by the respective studies. Figure 1 illustrates the classification of the current teacher population by their NBPTS status according to groupings used in the various studies.

The studies broke the population of teachers during each year of the study into seven total groups as shown in Figure 2. Group A consisted of current NBCTs. For a few analyses this group was subdivided into teachers in their first year of certification and other NBCTs. Group B comprised teachers who applied in the past and failed to receive certification. Group C included current applicants who will be certified within the timeframe of the study. Group D included current applicants who will fail to be certified or withdraw their application. Group E consisted of those who will apply and be certified in the span of the study but have not yet applied for certification. Group F was the failure counter part to Group E—i.e., it was the group of teachers who will fail to receive certification when they apply sometime after the current year. Finally Group G included the teachers who will have no involvement with NBPTS during the span of the study.

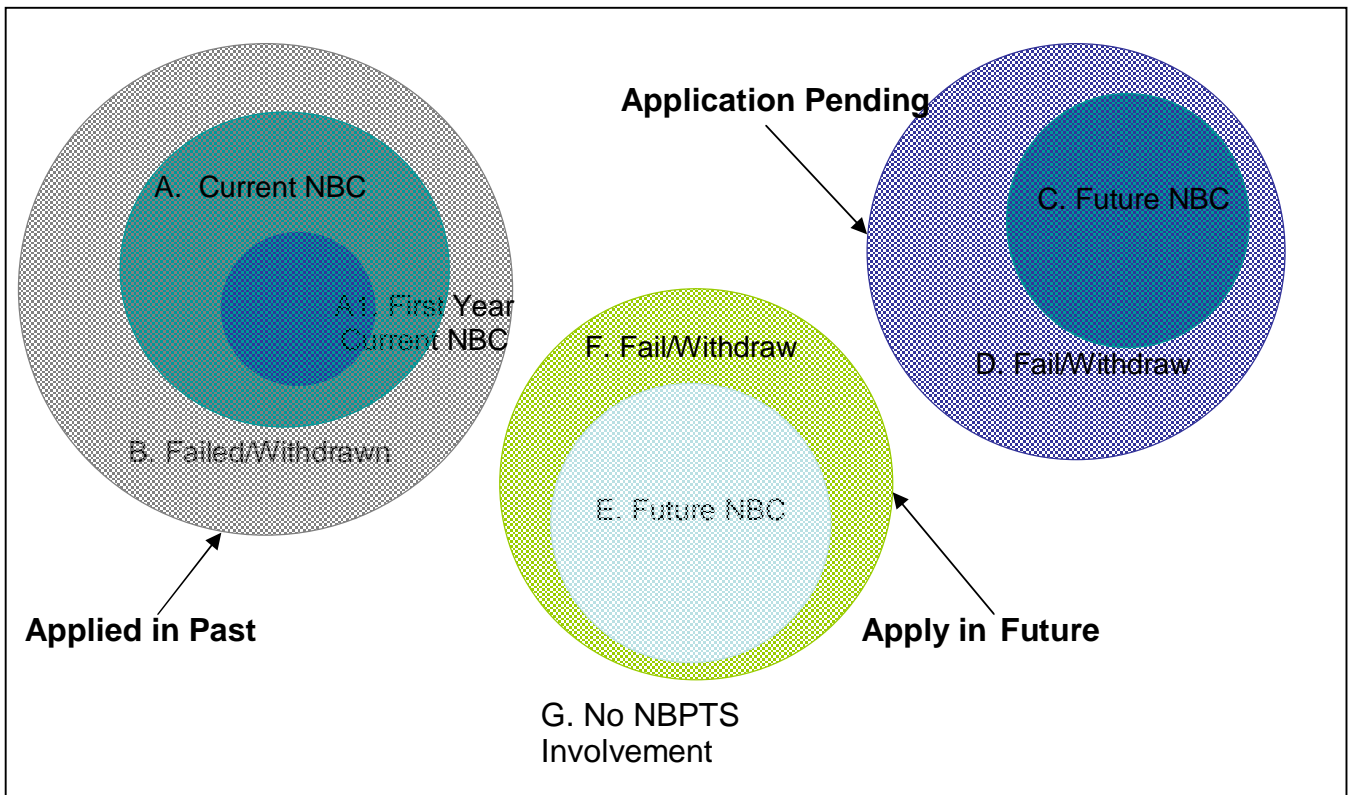


Figure 2. Diagram of Teacher Population by NBPTS Status

Table 1 provides an overview of the contrasts among the groups made by each study and the sign and significance of the test for the difference between groups (identified by the letters corresponding to the populations groups in Figure 2). For instance, Cavalluzzo (2004) fit models with indicator variables for current NBCTs (Group A), applicants with pending certification (Groups C and D) and teachers who applied for certification but failed to receive it (Group B). The holdout group, which was compared to each of the other groups, consisted of teachers who had not applied for certification, both those who would and those who would not apply for certification at a later time during the study period (Groups E and F and Group G). For those rows of Table 1, we show a comparison of Group A vs. the remainder group E+F+G, Groups C+D vs. E+F+G, and Group B vs. E+F+G. Cavalluzzo only considered mathematics so results for reading are not applicable. For A vs. E+F+G, the students of teachers in Group A scored significantly higher than students whose teachers were in E+F+G in all models, so there is a single green “+” in the column for mathematics. For students of teachers with pending applications (C+D) compared with students whose teachers had not applied (E+F+G), the results varied by model. In some models the differences were statistically significant and positive, but in other models the differences were negative but not significant. Hence, this row includes a green “+” to denote the finding of significant positive effects and a gray “-” to denote the finding of insignificant negative estimates. Significant negative differences are denoted by a red “-”.

One of most striking feature of Table 1 is the number of distinct contrasts estimated in these papers. As discussed above, the most policy relevant contrast is the comparison of current NBCTs to all other teachers. Only Clotfelter et al. (2006, forthcoming) make this comparison. In their 2006 paper, the achievement or gains in achievement of students of NBCTs was always greater than it is for similar students of other teachers. In their forthcoming paper, which is based on a subset of the data from the 2006 study, the differences were both negative and positive depending on the model. One possible explanation for the discrepancy between these results is that the estimates with the smaller sample are imprecise—the negative estimates were never significant—and the larger sample with substantially more teachers provides better estimates of the small but positive effect of NBCTs on students.

The studies by Goldhaber and Anthony (2005) and Harris and Sass (2006) both considered the contrast A vs. B+D+F+G, current NBCTs compared to all teachers who would never be certified during the span of the study. Goldhaber and Anthony (2005) found positive effects for NBCTs in this comparison for both reading and mathematics, although, only the reading effect was significant. Harris and Sass (2006) also found positive results for reading for both the FCAT-NRT and FCAT-SSS tests, although the difference was significant only for the FCAT-SSS. Harris and Sass found both positive and negative effects for mathematics but neither none were significant.

Although the comparison groups varied among the contrasts considered by the various studies, the differences typically involve teachers who had or would later apply for NBPTS certification. These teachers accounted for a small fraction of the total population of teachers and so their student might have had little effect on the means for the comparison group. If we ignore differences in the comparison groups (Table 2), we find that across all the studies nearly all the contrasts that involve current NBCTs yield positive effects for their students in reading (although they are not always significant).⁴ In fact, there were just three comparisons with negative effects and only one of those was significant. For mathematics, a majority of the models found positive effects for NBCTs. Again many of the differences were not statistically significant and in five instances a study found a negative effect for NBCTs but only one of these was significant. Thus, when comparing the students of current NBCTs to other students the studies provide somewhat consistent evidence that the achievement or achievement gains of the NBCTs' students was higher than that of the other students.

⁴ We exclude contrasts that involve only applicants or teachers who will apply for NBPTS certification.

Table 1. Summary of Population Subgroup Contrasts Estimated by Various Studies

Study	Population Subgroups		Reading	Mathematics
	Compared			
Cavalluzzo	A	vs. E+F+G	NA	+
	C+D	vs. E+F+G	NA	+, -
	B	vs. E+F+G	NA	-, -
Clotfelter, Ladd & Vigdor (2006) (Tables 2 & 3)	A	B+C+D+E+F+G	+, +	+
Clotfelter, Ladd & Vigdor (2006) (Table 6)	A-A1	vs. B+D+F+G	+	+
	A1	vs. B+D+F+G	+	+
	C	vs. B+D+F+G	+	+
	E	vs. B+D+F+G	+	+
Clotfelter, Ladd & Vigdor (forthcoming)	A	vs. B+C+D+E+F+G	+, +, -	+, -
Goldhabor & Anthony (1,2)	A	vs. B+D+F+G	+	+
	C+E	vs. B+D+F+G	+	+
	A-A1	vs. B+F+G	+	-
Goldhabor & Anthony (3)	A1	vs. B+F+G	+	+
	E	vs. B+F+G	+	+
	C+D	vs. B+F+G	-	-
Goldhabor & Anthony (4)	E+F	vs. G	+	+
	C+D	vs. G	-	-
	A+B	vs. G	+	-
	C+D	vs. B+F+G	-	-
Goldhabor & Anthony (5,6)	E	vs. B+F+G	+	+
	A-A1	vs. B+F+G	+, +	-, -
	A1	vs. B+F+G	+	+
	A+C+E	vs. B+D+F+G	-	-, +
Harris & Sass	A	vs. B+D+F+G	+, +	-, +
	C	vs. B+D+F+G	-, -	-, +
	E	vs. B+D+F+G	+, +	+, +
Sanders, Ashton & Wright	A	vs. G	+, +, -, -	+, +, -
	A	vs. C+D+E+F	+, -	+, +, -
	A	vs. B	+, +, -	+, +, -

Differences in Models Used by Various Studies

As noted above the studies also differed by the approaches taken to control for contribution of prior educational inputs and student background variables. The studies by Goldhaber and Anthony (2005) and Clotfelter et al. (2006) modeled achievement gains with extensive teacher, student and peer covariates. Both studies found positive effects for NBCTs with these models, although the effects for mathematics were not significant in Goldhaber and Anthony. Goldhaber and Anthony (2005) and Clotfelter et al. (2006) also considered models for gain scores that included either student or school fixed effects but not both. In both cases adding school fixed had no appreciable effect on the inferences about NBCTs. Adding student fixed effects had a very substantial impact on the estimates from Goldhaber and Anthony. For reading the positive and significant effects found in all other models became very small and not significant and for mathematics the results for current NBCTs past their first year as a certified teachers became very negative and statistically significant in the model with student fixed effects whereas it was insignificant, negative, but with a small magnitude in all the other models. However, the results in Clotfelter et al. (forthcoming) were almost invariant to the inclusion of student fixed effects but these models used achievement levels rather than gains as outcomes as were used by the Goldhaber and Anthony. Including fixed effects can dramatically change the sample used in estimating differences and some of the sensitivity found in the study Goldhaber and Anthony might be due to small samples of NBCTs used in that study and potential influential points or outliers that resulted when fixed effects were added to the model. Because of its larger sample size and, in particular, large number of teachers involved with the NBPTS, the study by Clotfelter et al. (2006) might have avoided some of these instabilities.

Overall, Clotfelter et al. (2006) provided a thorough comparison of the range of models generally considered among these papers. As noted above, their results are generally invariant to the model specification in term of the sign of the coefficient and the significance although the magnitude is somewhat sensitive to model specification. This might indicate that many of the differences we see among studies was more a result of using small samples of certified teachers to estimate very small effects for NBCTs (effect sizes of about .01 to .02) than it was a result of varying biases from the different adjustment procedures.

One model not considered by Clotfelter et al. (2006) is the model with fixed effects for schools and students that was used by Harris and Sass (2006). At the request of the panel, the authors did estimate these models, and the results are presented below.

Table 2. Summary of Contrast Involving Current NBCTs.

Study	Population Subgroups Compared			Reading	Mathematics
		vs.			
Cavalluzzo	A	vs.	E+F+G	NA	+
Clotfelter, Ladd & Vigdor (2006) (Tables 2 & 3)	A		B+C+D+E+F+G	+,+	+
Clotfelter, Ladd & Vigdor (2006) (Table 6)	A-A1	vs.	B+D+F+G	+	+
	A1	vs.	B+D+F+G	+	+
Clotfelter, Ladd & Vigdor (forthcoming)	A	vs.	B+C+D+E+F+G	+,+,-	+,-
Goldhaber & Anthony (1,2)	A	vs.	B+D+F+G	+	+
Goldhaber & Anthony (3)	A-A1	vs.	B+F+G	+	-
	A1	vs.	B+F+G	+	+
Goldhaber & Anthony (5, 6)	A-A1	vs.	B+F+G	+,+	-, -
	A1	vs.	B+F+G	+	+
Harris & Sass	A	vs.	B+D+F+G	+,+	-,+
Sanders, Ashton & Wright	A	vs.	G	+,+,-,-	+,+,-

Differences in the Populations Used by Various Studies

Of the studies with significant sample sizes, two used data from Florida (Cavalluzzo, 2004, and Harris and Sass, 2006) and the remainder use data from North Carolina (Goldhaber and Anthony, 2005, Clotfelter et al. 2006, forthcoming, and Sanders et al., 2005). Table 3 summarizes the populations used in each study.

Both studies using Florida data found positive effects for NBCTs relative to other grade 9 and 10 teachers when using the mathematics FCAT-SSS. Using level scores and covariate adjustment Cavalluzzo (2004) found large statistically significant effects for current NBCTs compared to teachers who had not applied for certification. Although Harris and Sass (2006) did not find significant results when comparing current NBCTs from all grades with teachers who were never certified during the span of their study, when they estimated effect separately for high school teachers, they found positive and statistically significant effects for current NBCTs on the FCAT-SSS.

There are also similarities among results from some of the studies using North Carolina data. For example, both Goldhaber and Anthony (2005) and Clotfelter et al. (forthcoming) report results from grade 5 teachers. In both cases the results were weak but positive for reading teachers and negative for mathematics teachers. However, using data from two of the state's largest school systems, Sanders et al. (2005) did not find a notable difference between mathematics and reading teachers. The Sanders et al. results were from a somewhat later time period and their models controlled for fewer variables and this could explain some of the difference. However, using 10 years of data for the entire state and grades 4 and 5, Clotfelter et al. (2006) found that the effects of NBCTs are almost twice as large for mathematics than for reading. Without additional analyses it is difficult to determine the source of this inconsistency.

Table 3. Summary of Populations Used in Each Study

Study	Outcome Test		Locations
	Years	Grades	
Cavalluzzo	2002 and 2003	9 and 10	Miami-Dade County
Harris and Sass	2001 to 2004	4 to 10	Florida
Sanders, Ashton & Wright	2001 to 2003	5 to 8	Charlotte-Mecklenberg and Wake County Schools
Goldhaber & Anthony	1998 and 1999	3 to 5	North Carolina
Clotfelter, Ladd & Vigdor (forthcoming)	2001	5	North Carolina
Clotfelter, Ladd & Vigdor (2006)	1996 to 2004	4 and 5*	North Carolina

* Some models include grade 3 outcomes.

Test Scores Used in the Analyses

One common feature to all these studies is that they use state accountability tests as the measure of student achievement. State accountability tests are of course of great interest to policy makers and educators, but they do not necessarily measure all aspects of achievement and they might not be most responsive to differences between the students of NBCTs and other teachers. For example, if NBCTs are better than other teachers at teaching higher-order skills not captured by basic proficiency tests or at teaching broad topics not covered by state tests, then the effects of NBCTs might not be captured by accountability tests; although they could be identified with alternative measures. In such a case, the results on the state test might provide a more negative picture of NBCTs than is true.

In addition state accountability tests typically have high stakes for educators and there is general concern about potential threats to the validity of high stakes tests for measuring

growth in achievement due to teaching to the test and other practices that can result in score inflation (Hamilton et al, 2005, Koretz et al. 2000). If teachers are engaging in such practices, then bias would be most likely if NBCTs tended to be more or less likely than other teachers to use practices that would distort test scores.

Although all studies reviewed by the report use high stakes tests, the study by Harris and Sass (2006) uses both the FCAT-SSS, the state's high stakes accountability test, and the FCAT-NRT, which is not used for determining adequate yearly progress or other high stakes decisions. Hence, we can use this study to investigate the sensitivity of results to the stakes associated with the test. Harris and Sass (2006) report considerable differences between results on the FCAT-NRT and the FCAT-SSS. Generally, the results were more positive on the FCAT-SSS than on the FCAT-NRT. It is difficult to determine the meaning of this difference. It could mean that the FCAT-SSS was more responsive to the differences between NBCTs and other teachers. It could mean that NBCTs were engaging in more practices that inflated scores in the high stakes test, than were other teachers. Alternatively the difference might mean that NBCTs were better at internalizing the state standards and teaching those standards because the FCAT-SSS is aligned with the standards. Additional research on the teaching practices of NBCTs is necessary to determine the actual source of the differences, but the existence of such difference suggests that the outcome measures can matter and that there may be more subtle differences between NBCTs and other teachers than can be determined by studies using only high stakes tests.

3. ADDITIONAL ANALYSES

In order to understand better the sources of the divergence in findings for Florida and North Carolina we asked Clotfelter, Ladd and Vigdor and Harris and Sass to estimate a series of mathematics and reading specifications over grades common to both samples. The authors graciously agreed to our request, and the results are reported in Tables 4 and 5. Table 4 uses student level data, and the specifications differ with respect to the parameterization of the NBPTS certification status variable (a single ever NBPTS certified variable, versus a three variable parameterization with future NBCT, year of or year following receipt of NBPTS certification, and became NBPTS at least two years prior), the structure of the value added model (gain as dependent variable or lagged achievement as an explanatory variable), and the type of fixed effects included to account for unobserved factors. Table 5 reports specifications based on data aggregated to the school by grade by year level in order to avoid biases introduced by the non-random sorting of students among classrooms. These specifications also vary along the same dimensions as those reported in Table 4.

The results in Table 4 continue to show a sharp divergence by state in NBPTS effects on mathematics, but the effects for reading are quite similar though a bit smaller and less precisely estimated in Florida. Specifically, using the simple ever certified specifications (model 2), the estimated effect sizes are somewhere between 50 and 100 percent larger in North Carolina depending upon the specification. Interestingly, there is far less sensitivity to specification in Florida. A point of difference in the reading results is the pattern of

estimates in Model 1 that permits variation by timing with regard to NBPTS status. The results for Florida suggest that the application process is costly in terms of reducing quality and that teachers learn little from the process (certified in the future and certified in prior year coefficients are almost identical), while the results for North Carolina are much more uncertain regarding both costs and learning. In all specifications the certification differential is larger for already certified teachers, and the coefficient for those currently obtaining certification fluctuates quite a bit though always remaining below the coefficient for those already certified.

Similar to the findings reported by the authors in their papers, the average NBCT effects in mathematics are larger than in reading in North Carolina but not significantly different from zero in any specification with fixed effects in Florida. In addition, there is some evidence of learning in North Carolina but also some evidence that the process is costly.

Because of the possibility that non-random allocation of students into classrooms influences estimates of NBCT effects, we also asked the authors to produce estimates based on data aggregated to the school by grade by year level. In this framework the NBCT effects are identified by variations across grades and years in the share of ever NBCTs (model 2) or shares of teachers in different NBPTS stages. To control for unobserved differences among students and teachers across schools, models include school by grade and/or school by year fixed effects in the regressions reported in Table 5. Not only do the school by year effects account for all between school differences, but they also account for any shock to the school or neighborhood that may be correlated with the share of NBCTs. In addition, the school by grade fixed effects account for any grade specific experiences in a school that are correlated with the probability of a student having an NBCT.

Unfortunately the North Carolina authors were not able to provide the aggregate estimates as of this writing, but the Florida results reported in Table 5 generally confirm the finding of no NBCT effects in mathematics and, if anything, suggest that the effects of NBCTs are more positive in reading than is indicated by the student level analysis. The coefficients on ever certified in the full fixed effects models are roughly five times as large as the corresponding coefficients in the student level specifications. Although error in classifying teachers as NBCTs might contribute to this difference, a much more compelling explanation is that principals tend to assign more difficult to educate classes to NBCTs, and that such student heterogeneity is not captured adequately in the student level models. The fact that the aggregate models are sensitive to specification suggests that time and grade varying confounding factors might have important influences.

5. SUMMARY

In response to growing concerns about the quality of the primary and secondary education and the supply of teachers, the National Board of Professional Teacher Standards developed standards for what accomplished teachers should know and be able to do and criteria procedures for certifying who meet those standards. In the past few years, several studies, many with full or partial funding from the NBPTS, estimated the

effect of NBPTS certified teachers on their students' achievement or achievement growth. This report reviewed eight of these studies in some detail and also summarized the results of new analyses on two large student test score databases that were commissioned by the NRC.

All the studies, except one with a very small sample, used data from Florida or North Carolina, but they considered a vast array of model specifications to account for the nonrandom assignment of students to teachers' classes and to distinguish between the signaling and learning effects of NBPTS certification. Across this wide array of analyses, the estimates of the effects NBPTS certification vary from positive and statistically significant to negative and statistically significant.

One source of variance in the estimated effects is the differences in the groups of teachers being compared by the estimators of the NBCT effects. The studies classified teachers into several strata on the basis of their involvement with NPBTS. Different studies contrasted different strata in various attempts distinguish between the signaling and the learning effects of the certification process or to differentiate applicants who did or did not receive certification. However, across studies, the estimates of the effect of current NBCTs compared to other teachers present a consistent pattern of positive effects in reading; although in many studies the effects are not significant. For mathematics there are also more positive effects than negative effects, four of the six studies find negative effects for current NBCTs on mathematics achievement for at least some of the model specifications they used. Again many estimates are not significant. The most comprehensive study conducted by Clotfelter et al. (2006) used 10 years of data from North Carolina and found consistent positive effects for reading and mathematics for fourth and fifth grade teachers and students. However, the effects are very small with effect sizes of about .01 and .02 standard deviations units for reading and mathematics respectively. These estimates were relatively insensitive to model specification.

To isolate the effects of different model specifications on the estimates, Harris and Sass and Clotfelter, Ladd and Vigdor, fit a large suite of additional models to data from Florida and North Carolina. Both groups used data from the same grades and years and fit exactly the same model specifications. The models specifications were determined by a series of five factors for the specification of the outcome (achievement level or gain), the use of student fixed effects, the use of school fixed effects, the inclusion of teacher characteristics, and the specification of the NBCT effect. Models were fit using both individual student data and data aggregated to the school by grade by year level.

The results of this unique head-to-head comparison suggest results can be more sensitive to context than model specification. Across the models considered, there are consistent positive and significant effects for NBCTs in North Carolina. This finding holds for both reading and mathematics. The magnitude of the effects varies with model specification but the sign and significance do not. In Florida, the effects of NBCTs on reading achievement are smaller than in North Carolina, but they are always positive and significant for all but one model. For mathematics the Florida estimates are very small and insignificant.

Thus, the results seem to be sensitive to subject and state but much less so to model specification. Hence, we conclude that NBCTs have positive effects on student achievement and gains in achievement in mathematics and reading for fourth and fifth graders in North Carolina as measured by that state's accountability test. The effects of NBCTs on the achievement of Florida fourth and fifth grader as measured by FCAT-SSS are positive for reading, but than in North Carolina, and very small and indistinguishable from zero for mathematics.

The source of the differences across states and subjects remains an important topic for future research. It might be related to the nature of the tests. Estimates in Florida are sensitive to the specific test with estimates based on the low-stakes FCAT-NRT smaller than those based on the FCAT-SSS (as reported in Tables 4 and 5). Difference between the states might also depend on unmeasured characteristics of teachers who apply for NBPTS certification in each state. It could also depend on subtle factors such as the relationship between NBPTS standards and state standards. NBPTS have a long history of support in North Carolina and this could have an impact on the teachers who apply and how they perform relative to the state standards. Additional studies that probe these and other hypotheses would be very useful for providing greater understanding of meaning of NBCTs for student achievement.

The effect of context on the estimated effect of NBCTs is not limited to difference by state or subject; grade level appears to matter as well. The Florida database allows for estimation of effects for elementary, middle school, and high school teachers. As discussed above, both Cavalluzzo (2004) and Harris and Sass (2006) found significant effects for NBCTs on ninth and tenth grade students' mathematics achievement and achievement gains as measured by the FCAT-SSS. Harris and Sass also ran the additional suite of models on the high school data from Florida and found consistent positive effects for both the FCAT-SSS and the FCAT-NRT. This suggests that the effects of NBCTs on mathematics might be larger in high school than in elementary school. This finding is consistent with other studies that find that certain teacher qualifications had greater effects on high school students where teachers' content knowledge might be particularly important. This interesting finding needs further investigation to confirm that the findings in Florida are not spurious and to determine if it replicates to other contexts.

This very rich literature on the effects of NBCTs on student achievement provides some evidence of small positive effects. When the sample sizes are very large, the findings tend to be robust to model specification, suggesting that much of the variability of estimates found in other studies was most likely the effects of sampling error due to the small samples of NBCTs used in those studies. However, there is evidence that these effects can be sensitive to context, subject and grade-level and the future of research on NBCTs should focus on understanding the sources of these differences.

Table 4. Estimated Effects of National Board Certification on Mathematics and Reading Scores in Florida and North Carolina from Student Level Regressions (standard errors in parentheses)

Model	Mathematics						Reading					
	gain	gain	gain	gain	lagged score	lagged score	gain	gain	gain	gain	lagged score	lagged score
student fixed effects	no	yes	no	yes	no	no	no	yes	no	yes	no	no
school fixed effects	no	no	yes	yes	no	yes	no	no	yes	yes	no	yes
<i>Florida</i>												
Model 1												
certified in the future	0.004 (0.012)	-0.007 (0.015)	-0.002 (0.012)	-0.010 (0.015)	0.013 (0.011)	-0.002 (0.011)	0.019 (0.010)	0.046 (0.014)	0.021 (0.010)	0.043 (0.015)	0.020 (0.010)	0.016 (0.010)
certified in current year	-0.006 (0.013)	-0.005 (0.017)	-0.016 (0.013)	0.005 (0.018)	0.005 (0.013)	-0.011 (0.012)	0.005 (0.011)	-0.010 (0.016)	0.008 (0.011)	-0.009 (0.017)	0.007 (0.011)	0.007 (0.010)
certified in prior year	0.013 (0.009)	0.011 (0.011)	0.013 (0.009)	0.022 (0.011)	0.025 (0.008)	0.018 (0.008)	0.010 (0.007)	0.009 (0.009)	0.013 (0.007)	0.009 (0.010)	0.018 (0.006)	0.019 (0.007)
Model 2												
ever certified	0.003 (0.008)	0.007 (0.006)	0.003 (0.007)	0.011 (0.009)	0.018 (0.006)	0.007 (0.006)	0.011 (0.005)	0.015 (0.007)	0.014 (0.006)	0.015 (0.008)	0.016 (0.005)	0.016 (0.005)
<i>North Carolina</i>												
Model 1												
certified in the future	0.049 (0.011)	0.058 (0.015)	0.042 (0.011)	0.077 (0.016)	0.044 (0.011)	0.035 (0.010)	0.024 (0.009)	0.037 (0.012)	0.020 (0.009)	0.047 (0.012)	0.021 (0.009)	0.013 (0.008)
certified in current year	0.027 (0.013)	0.044 (0.014)	0.029 (0.013)	0.045 (0.015)	0.033 (0.012)	0.033 (0.012)	0.021 (0.009)	0.024 (0.012)	0.016 (0.009)	0.015 (0.013)	0.029 (0.009)	0.022 (0.009)
certified in prior year	0.055 (0.006)	0.074 (0.007)	0.051 (0.006)	0.078 (0.008)	0.062 (0.006)	0.053 (0.006)	0.027 (0.004)	0.042 (0.006)	0.024 (0.005)	0.038 (0.007)	0.035 (0.004)	0.027 (0.004)
Model 2												
ever certified	0.050 (0.005)	0.067 (0.006)	0.046 (0.005)	0.072 (0.007)	0.055 (0.005)	0.047 (0.005)	0.026 (0.004)	0.038 (0.005)	0.022 (0.004)	0.036 (0.006)	0.032 (0.004)	0.024 (0.004)

Table 5. Estimated Effects of National Board Certification on Mathematics and Reading Scores in Florida and North Carolina from School by Grade Level Regressions (standard errors in parentheses; regressions weighted by cell size)

Model	Mathematics						Reading					
	gain	gain	gain	lagged score	lagged score	lagged score	gain	gain	gain	lagged score	lagged score	lagged score
school by grade f.e.	yes	no	yes	yes	no	yes	yes	no	yes	yes	no	yes
school by year f.e.	no	yes	yes	no	yes	yes	no	yes	yes	no	yes	yes
<i>Florida</i>												
Model 1												
certified in the future	0.152 (0.067)	-0.047 (0.080)	-0.055 (0.075)	-0.028 (0.057)	-0.061 (0.039)	-0.108 (0.075)	0.375 (0.078)	0.146 (0.072)	0.033 (0.070)	0.024 (0.036)	0.062 (0.026)	0.073 (0.055)
certified in current year	0.028 (0.062)	-0.071 (0.099)	-0.064 (0.068)	0.016 (0.052)	-0.041 (0.048)	-0.009 (0.067)	-0.066 (0.077)	-0.155 (0.090)	0.039 (0.065)	0.048 (0.035)	0.048 (0.032)	0.089 (0.051)
certified in prior year	0.051 (0.049)	-0.077 (0.068)	0.060 (0.054)	0.075 (0.042)	0.049 (0.033)	0.078 (0.054)	-0.160 (0.065)	0.047 (0.063)	0.106 (0.052)	0.025 (0.030)	0.026 (0.022)	0.071 (0.041)
Model 2												
ever certified	0.068 (0.043)	-0.066 (0.047)	0.013 (0.046)	0.041 (0.036)	-0.006 (0.023)	0.023 (0.045)	0.000 (0.054)	0.037 (0.043)	0.076 (0.044)	0.028 (0.025)	0.042 (0.015)	0.075 (0.034)

REFERENCES

Carnegie Forum on Education and the Economy. (1986). *A nation prepared: Teachers for the 21st century, the report of the Task Force on Teaching as a Profession*. Washington, DC: Author.

Cavalluzzo, L. C. (2004). *Is National Board certification an effective signal of teacher quality?* Alexandria, VA: The CNA Corporation.

Clotfelter, C. T., Ladd, H. F. and Vigdor, J. L. (2006). How and why do teacher credentials matter for student achievement? Unpublished manuscript.

Clotfelter, C. T., Ladd, H. F. and Vigdor, J. L. (Forthcoming). Teacher sorting, teacher shopping, and the assessment of teacher effectiveness. *Journal of Human Resources*.

Goldhaber, D. and Anthony, E. (2005). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. Unpublished manuscript.

Hamilton, L. S., McCaffrey, D. F., and Koretz, D. M. (2006) Validating achievement gains in cohort-to-cohort and individual growth-based modeling contexts” in R. Lissitz (Ed.), *Longitudinal and Value-Added Modeling of Student Performance*, (pp. 407-434). Maple Grove, MN: JAM Press.

Harris, D. N. and Sass, T. R. (2006). The effects of NBPTS-certified teachers on student achievement. Unpublished manuscript.

Koretz, D. M., McCaffrey, D. F., and Hamilton, L. S. (2001). *Toward a framework for validating gains under high-stakes conditions* (CSE Technical Report 551). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

McColsky, W., Stronge, J. H., Ward, T. J., Tucker, P., Howard, B., Lewis, K. Hindman, J. L. (2005). *Teacher effectiveness, student achievement, & National Board certified teachers: A Comparison of National Board certified teachers and non-National Board certified teachers: Is there a difference in teacher effectiveness and student achievement?* Unpublished report to the National Board of Professional Teachers Standards. Retrieved February 22, 2007 from <http://www.wm.edu/education/Teacher%20Effectiveness.pdf>.

Sanders, W. L., Ashton, J. J., and Wright, S. P. (2005). *Comparison of the effects of NBPTS certified teachers with other teachers on the rate of student academic progress*. Unpublished report to the National Board of Professional Teachers Standards.

Spence, A. M. (1973) Job market signaling. *Quarterly Journal of Economics* 87(3), 355-74.

Vandervoort, L. G., Amrein-Beardsley, A. and Berliner, D.C. (2004). National Board certified teachers and their students achievement. *Education Policy Analysis Archives* 12(46). Retrieved July, 2006 from <http://epaa.asu.edu/epaa/v12n46/>.