Research Report

The Impact of the National Board for Professional Teaching Standards: A Review of the Research

Drew Gitomer

Research & Development

July 2007
RR-07-33

The Impact of the National Board for Professional Teaching Standards: A Review of the Research

Drew H. Gitomer
ETS, Princeton, NJ

July 2007

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

Interest in the impact of the teacher advanced certification system instituted by the National Board for Professional Teaching Standards (NBPTS) has resulted in a quickly expanding body of literature. This review examines five critical issues covered in the literature: the extent to which National Board certified teachers (NBCTs) engage in stronger instructional practice, are associated with stronger learning gains in their classrooms and schools, benefit from participating in the assessment process, influence the teaching quality of their peers, and are present in schools and classrooms with different characteristics. While most comparisons provide evidence that support the incremental effectiveness of NBCTs, the effects are generally modest. However, all findings must be interpreted with caution, given the limitations of research designs, methods, and tools used to explore teacher effects.

Key words: National Board for Professional Teaching Standards, teacher certification, teacher quality, teacher effectiveness, instructional practice, learning gains

Acknowledgments

By way of full disclosure, the author of this document is a researcher at ETS, the general contractor for the NBPTS assessment. The author has worked directly with NBPTS in a variety of ETS roles for almost a decade. Despite a conscious attempt to be objective in this report, this potential conflict of interest is acknowledged.

I would like to thank the National Board for Professional Teaching Standards for supporting this work. Leslie Stickler was an invaluable help in conducting the literature review as well as reviewing the work as it developed. I very much appreciate the thoughtful and helpful reviews of Joan Auchter, Henry Braun, Mary Dilworth, Carol Dwyer, Shannon Fox, Judy Koenig, Mari Pearlman, Joan Snowden, Cindy Tocci, and the NBPTS ACAP committee. I am also grateful to Jane Cairns for support in the production of the review. Of course, all information presented is solely the responsibility of the author.

Table of Contents

Introduction

As the National Board for Professional Teaching Standards (NBPTS) reaches its 20th birthday, it continues to be a focus of great educational and policy interest. Founded in 1987 as an outgrowth of *A Nation Prepared: Teachers for the 21st Century* by the Carnegie Task Force on Teaching as a Profession (Carnegie Corporation of New York, 1986), NBPTS has articulated an extremely ambitious set of goals for improving the practice of teaching and American education itself by:

- Maintaining high and rigorous standards for what accomplished teachers should know and be able to do.
- Providing a national voluntary system certifying teachers who meet these standards.
- Advocating related education reforms to integrate National Board Certification (NBC) in American education and to capitalize on the expertise of National Board certified teachers (NBCTs).

Significant dollars have been expended for NBPTS certification. Expenses include costs to apply and take the assessment, support for preparation efforts, salary bonuses and increments, and research. The most significant sources of support have been state and federal grants and contributions from private foundations. All told, approximately $438 million in federal and nonfederal support have been contributed to NBPTS efforts between 1987 and 2006. Approximately one third of those expenditures ($145 million) have been federal dollars. Since 2004, the proportion of federal contributions has been closer to 25% per annum.[1]

Such an unprecedented effort to improve the quality of teaching raises the critical question of the extent to which NBPTS is actually achieving the ambitious set of goals it has championed. During the last 5 years, a significant number of research studies have been conducted to address particular aspects of the impact of NBPTS certification. The purpose of this report is to review this body of research and to identify important findings and issues raised. A much more comprehensive study about NBPTS is currently being conducted by the National Academy of Sciences (NAS). The NAS report is expected to be published in late 2007. All research syntheses provide unique perspectives on the meaning of a body of research. It is hoped that, as with all aspects of inquiry in the sciences and social sciences, this perspective will provide incremental value to other ongoing efforts.

As will become evident from this review, the body of research is impressive, comprehensive, and far-reaching, but is far from conclusive. Indeed, in the search for answers about the effectiveness of NBPTS, some claims have been overstated by either the authors or other interested parties. The intent of this paper is to provide an objective analysis of the findings.

Research Questions Addressed in This Review

From its inception, NBPTS has sponsored an extensive research program. The primary focus during the first decade was on psychometric considerations involved in the design and scoring of, arguably, the most complex large-scale performance assessment in education ever developed. This research, led by the NBPTS Technical Advisory Group, chaired by Richard Jaeger, focused on establishing the basis for an assessment that satisfied professional guidelines of testing practice, with particular attention to issues of validity, reliability, and fairness. A list of these studies is available at the NBPTS Web site (www.nbpts.org/resources/research).

As the assessment became stable and the pool of candidates to explore particular questions grew, the research began to address issues of external validity and efficacy. The issue of NBPTS certification effectiveness is multidimensional and involves a set of subsidiary questions. Five core questions have dominated this research and are the focus of this review:

1. *To what extent does the NBPTS certification process identify strong instructional practice?* The NBPTS assessment process purports to identify highly accomplished teachers. Therefore, teachers identified by this process ought to represent the qualities articulated in the NBPTS standards (National Board of Professional Teacher Standards, n.d.), which describe accomplished teaching in respective fields What is the status of the external validity evidence to support the claim that NBCTs are identifiably different from their peers in terms of their instructional practices?

2. *To what extent does the NBPTS certification process identify teachers who differ in terms of their students' academic achievement?* It is insufficient to define effective teaching only in terms of the instructional actions of the teacher. Ultimately, effectiveness must be considered in terms of teachers' impact on students. Do students of NBCTs show different patterns of achievement than students of non-NBCTs?

3. *To what extent does participation in the NBPTS certification process enhance teacher effectiveness?* A key intention underlying the design of the NBPTS assessment process is to develop an assessment that, in itself, is a unique professional development effort to improve teacher effectiveness, independent of the outcome of the assessment. By having teachers consider the standards and present and analyze their teaching, NBPTS aims to engender reflective practice (Schön, 1987) that will facilitate effective teaching. Thus, research has been directed at examining the extent to which teacher effectiveness is affected by participation as National Board candidates.

4. *To what extent are schools more effective due to the influence of NBCTs on the practice of their non-NBCT colleagues?* Another key intention of NBPTS is that the certification process identifies and acknowledges teachers who are leaders in their profession and local school communities. Indeed, part of the assessment involves documenting prior professional leadership roles of the candidate. Through mentoring, leadership, and general influence, does the teaching practice of the non-NBCT faculty benefit from the presence of NBCTs in the school?

5. *To what extent does NBPTS contribute to the equitable distribution of teachers?* Traditionally, schools and children facing the most challenging circumstances also have the least access to qualified, let alone highly accomplished teachers (e.g., Clotfelter, Ladd, & Vigdor, 2004; Ingersoll, 2004). Through professional development and support efforts, another goal of NBPTS has been to equalize access to high-quality teachers for all students, regardless of economic and geographic circumstances. Indeed, the intellectual founders of NBPTS felt that states and districts would implement policies to improve the distribution of high-quality (NBPTS) teachers in high-need schools (Koppich, Humphrey, & Hough, 2007).

There are several important points to be made about the research conducted on NBPTS. First, NBPTS has been the largest supporter of this work, providing funding[2] and/or data access to many researchers. Second, research on the NBPTS program must be viewed in the context of significant public debate and reform efforts surrounding requisite knowledge for teaching,

teacher education programs, curriculum, compensation, teacher regulation, unionization, and the federal government's role in education. While this review cannot begin to pursue these issues, it is of some relevance that the political interest in claims made about teaching on the basis of NBPTS research often extends beyond the specific boundaries of the NBPTS program.

Scope and Limitations of This Review

This review focuses only on studies that address the aforementioned questions. Studies related to the psychometric defensibility of the assessment are not addressed. This review included most, but not all, the research that has been written to date on the respective topics. Because this is such an emerging literature, many of the research papers have not yet been published in peer-reviewed journals. Therefore, for each question, the author made a judgment as to the interpretability of results, and the standard was likely more lenient than would be used in a field with a more established research base.

While almost every study has significant sampling limitations, often acknowledged by the authors, such limitations suggest significant caution to the universe to which the results should be applied. However, because these studies tell us something about NBPTS impact, even if limited to a greater extent than suggested by authors' conclusions, these studies were included. However, there were a small number of studies that were uninterpretable because of severe methodological inadequacies and almost all were excluded from the review, even after applying a liberal standard. The exceptions were those studies that have received significant attention and been involved in public policy discussions despite their problematic nature.

Many of these studies involve complex methodologies, particularly those involved in determining the value-added contributions of teachers to student performance. The adequacy of these methodologies is hardly a settled issue and is the subject of continued research and debate that goes beyond evaluations of the effectiveness of NBPTS certification. While this paper does not fully engage in evaluating these measurement challenges, it tries to acknowledge, as appropriate, issues relevant to the interpretation of findings.

The appendix contains a brief summary of the studies addressed in this review.

A Brief Primer on NBPTS Assessment

To make sense of NBPTS research, it is important to understand the structure of the assessment. Detailed descriptions of the assessments for each teaching field are available on the

NBPTS Web site (http://www.nbpts.org/the_standards/standards_by_cert), all of which adhere to the same general structure.

Assessments are offered for 24 teaching fields that vary according to discipline (e.g., Mathematics, English Language Arts, Music) and students' age (e.g., Early Childhood, Middle Childhood, Early Adolescence). The number of teachers in each field, and consequently, the number of teachers applying for certification in each field, varies significantly. A listing of the certifications, number of first-time candidates in 2005–2006, and total number of candidates certified through 2006 is presented in Table 1.

Table 1

Summary of NBPTS Certificates

Certificate	Abbreviations	Age range of students	Number of candidates 2005/2006	Number of candidates certified 1993–2006
Early Childhood/Generalist	EC/Gen	3–8	3,324	9,695
Middle Childhood/Generalist	MC/Gen	7–12	3,059	9,448
Early Adolescence through Young Adulthood/Art	EAYA/Art	11–18+	258	1,118
Early Adolescence/ English Language Arts	EA/ELA	11–15	965	2,999
Adolescence and Young Adulthood/Math	AYA/Math	14–18+	537	2,087
Adolescence and Young Adulthood/Science	AYA/Science	14–18+	742	3,098
Adolescence and Young Adulthood/English Language Arts	AYA/ELA	14–18+	1,005	3,334
Early Adolescence/Math	EA/Math	11–15	810	2,820
Early Adolescence/Science	EA/Science	11–15	618	1,616
Early Adolescence/ Social Studies-History	EA/SSH	11–15	349	1,200
Adolescence and Young Adulthood/Social Studies-History	AYA/SSH	14–18+	562	1717
Early and Middle Childhood/ English as a New Language	EMC/ENA	3–12	181	784

(Table continues)

Table 1 (continued)

Certificate	Abbreviations	Age range of students	Number of candidates 2005/2006	Number of candidates certified 1993–2006
Early Adolescence through Young Adulthood/ English as a New Language	EAYA/ENA	11–18+	95	228
Early Childhood through Young Adulthood/ Exceptional Needs Specialist	ECYA/ENS	birth–21+	1,759	3,929
Early Adolescence through Young Adulthood/ Career and Technical Education	EAYA/CTE	11–18+	695	2,285
Early and Middle Childhood/Literacy: Reading-Language Arts	EMC/LRLA	3–12	211	1450
Early and Middle Childhood/Physical Education	EMC/PE	3–12	162	526
Early Adolescence through Young Adulthood/Physical Education	EAYA/PE	11–18+	209	436
Early and Middle Childhood/Art	EMC/Art	3–12	458	486
Early Childhood through Young Adulthood/Library Media	ECYA/LM	3–18+	266	1,619
Early Childhood through Young Adulthood/School Counseling	ECYA/SC	Pre-K-12th grade	245	943
Early and Middle Childhood/Music	EMC/Music	3–12	308	730
Early Adolescence through Young Adulthood/Music	EAYA/Music	11–18+	698	520
Early Adolescence through Young Adulthood/World Languages Other than English	EAYA/WLOE	11–18+	1,616	753
Totals			19,132	53,821

The assessment contains two major components. The *portfolio* consists of a set of entries in which teachers provide evidence of their practice. These entries contain classroom artifacts such as videotapes of lessons and instructional materials and a detailed written analysis of the

teachers' practice. For each entry, teachers must present particular aspects of their practice central to the respective standards of their teaching field. Specific expectations for portfolio entries differ across fields, while the general entry structures, including presentation of artifacts and expectations for written analysis, remain consistent. The portfolio also asks teachers to provide evidence of accomplishments outside the classroom, including professional activities and involvement with students' families.

The second segment of the assessment is known as the *assessment center* and involves a set of constructed-response tasks that candidates take in a secure environment. These tasks are designed to assess subject matter knowledge important to their field of teaching, as articulated in the NBPTS standards.

The number of assessment tasks in each segment was changed in an assessment redesign implemented in 2001. However, the basic weighting of the two segments has not changed. The portfolio accounts for 60% of a candidate's total score and the assessment center accounts for the remaining 40%.[3] All tasks are scored according to an elaborate process that has also evolved over years of experience with the NBPTS certification process.

Introduction to Findings

This review was undertaken to determine whether any broad and general claims could be made about NBPTS and its impact on teachers, students, and schools. The answer, detailed in the following sections, is that conclusions must be viewed as extremely tentative. This uncertainty applies to studies offering strong support for NBPTS as well as to those raising serious doubts. The tentativeness of findings is most readily ascribed to methodological limitations of the studies that have been undertaken. In most cases, the critiques are not the function of any lapses of researchers, but rather result from fundamental limits on the availability of data and appropriate methodologies. While this paper raises specific concerns in the discussions of many of these studies, a number of recurrent issues are summarized at the outset.

The essential research problem involves comparing NBCTs with other teachers on a set of criterion measures. Executing such comparisons turns out to be a relatively complicated task. Specific limitations to the research include the following:

1. *Selecting the appropriate comparison group.* The issue of selecting an appropriate comparison group is an important consideration in all NBPTS research. In any

evaluation of the effectiveness of NBPTS, the question must be raised: Effective compared to whom? If one asks whether the NBCT assessment signals a certain level of effectiveness, then simply comparing NBCTs to other non-NBCT teachers is problematic. The latter group is likely to include many teachers who would qualify as NBCTs had they chosen to participate in the NBPTS process. Because there is no way of knowing the nature of this comparison group, making inferences about relative effectiveness is much more ambiguous and certainly has different implications than for comparing candidates who succeeded and failed in attaining certification. The reason that many studies do not compare successful and unsuccessful candidates is because of the availability of relevant data for sufficiently large numbers of unsuccessful candidates.

2. *Representativeness of NBCTs and comparison groups.* To generalize from a research sample to a broader population, the sample must be representative of the population. The optimal way to achieve this is to have a random sample of the population. That is, it is desirous to have a random sample of NBCTs and a random sample of teachers in the comparison group of interest. Short of randomization, one can try to approximate the distribution of characteristics of the population through sample selection (e.g., race, age, socioeconomic status [SES] of schools). However, for many NBPTS studies, logistical and practical factors, including self-selection of participants (both NBCTs and non-NBCTs), constrained the sample design. This fact makes generalization beyond the sample much more tenuous.

3. *Controlling for student and other factors.* Conclusions based on comparing two groups of teachers on some measures of interest assumes that any differences can be ascribed to teacher differences rather than any confounding variables. As will become abundantly clear, such assumptions are inappropriate. Teachers are not randomly assigned to schools and students are not randomly assigned to teachers. While there are ways to imagine a randomized experiment that would assign teachers to different classrooms, or perhaps multiple classrooms over several years, such a study seems practically infeasible. Therefore, some of the research attempts to statistically control for these factors, while other studies do not. But, even when controls are used, methodological questions remain and will be discussed further.

4. *Sample size and statistical power* Even with the relatively large recent number of NBPTS candidates (see Table 1), the actual number of teachers available for study is often quite small. Most studies look at one or two certification areas only, reducing the available pool substantially. More important, however, any comparisons with state-related data (e.g., student assessment results) mean that analyses can only be conducted within-state. Once the constraints of certification area and state are imposed, along with participation and availability of data, the actual sample sizes of NBCTs and relevant comparison groups tend to be small. The consequence is that statistical power is low, making the observation of differences, even if they actually exist, much more elusive.

   A related sampling issue is that participation in NBPTS has varied dramatically across states. North Carolina has adopted policies to encourage NBPTS participation from its inception and Governor James Hunt of North Carolina was the founding chair of NBPTS. Thus, an overwhelming majority of studies have been conducted in North Carolina. Interpretations based only on data from one or two states must be viewed with appropriate caution.

   A final sampling issue is that, due to availability of data, most studies have been conducted with the Early and Middle Childhood Generalist certifications, the areas with the largest NBCT volumes. The ability to make claims about teachers in specific disciplines on the basis of generalist results is undetermined.

5. *Emerging methodologies.* Popular, relatively new methods, such as value-added models (VAMs), are being used widely to explore teacher effects. Yet, the understanding of these methods and the assessment measures that are used as the data for these methods, are still relatively immature. These limitations will be explored further in the section examining student achievement effects.

With these caveats, I next review studies addressing each of the five major questions. In each section I review results and also address specific methodological limitations of the studies, individually and collectively. The review closes with a summary of the status of the research on NBPTS and recommendations for next steps.

Studies of Instructional Practice

A large body of research literature exists on effective teaching practices. Many of these studies are based on relatively small samples, but syntheses have made clear that certain instructional practices are associated with improved student learning. Such practices include questioning (e.g., Redfield & Rousseau, 1981), assessment and feedback (Black & Wiliam, 1998), and establishing important content-relevant goals (Bransford, Brown, & Cocking, 1999). The findings of many of these studies are articulated in the NBPTS Standards.

The studies reviewed in this section focus on the instructional practices of NBCTs compared to peers. These studies are predicated on the inference that the use of teaching practices identified as effective will lead to more successful student learning. Each study identifies particular instructional practices and examines the extent to which NBCTs use these instructional practices with different frequency than do other teachers. These studies, with a few exceptions, do not include direct measures of student achievement.

The first major study of instructional practices of NBCTs examined 65 teachers from North Carolina who applied for certification in the fields of Early Adolescence/English Language Arts and Middle Childhood/Generalist (Bond, Smith, Baker, & Hattie, 2000). The authors compared the performance of NBCTs with those who applied for certification but were not successful. Teachers from both certification fields were combined in the analyses. Using a variety of data sources, they examined evidence about differential performance on 13 dimensions of teaching that have been established in the general literature as effective teaching practice.

A virtue of the Bond et al. (2000) study is that there is a comparison between known groups—individuals who were successful on the assessment compared to those who were not. However, the analysis does not account for the possibility that instructional performance might also be a function of the students taught and might differ across classrooms. Bond et al. did attempt to structure their sample by matching NBCT and non-NBCT teachers on a variety of comparison measures, but admittedly, they were not entirely successful.

With these methodological issues in mind, I turn to the study itself. The authors identified 13 dimensions of teaching, presented in Table 2, that have been associated in the research literature with teacher expertise and effectiveness. These dimensions grew out of an emerging body of research on expertise in teaching and distilled in frameworks by Shulman (1987) and Sternberg and Horvath (1995). Dimensions include aspects of conceptual rigor and depth (e.g.,

deep representations and problem solving in Table 2) as coded using the Structure of Learning Objectives (SOLO) taxonomy (Biggs & Collis, 1982), pedagogical expertise (e.g., monitoring learning and providing feedback), as well as interpersonal characteristics (e.g., respect for students). They then collected multiple measures of teacher practice, including lesson transcripts from observations, preobservation questions, teacher interviews, interviews of randomly selected students from the teacher's classroom, assignment logs, student questionnaires, student work samples, and student writing samples. For most dimensions, multiple evidence sources contributed to the evaluation. All were coded by trained judges, blind to the certification status of the teachers.

Table 2

Dimensions of Expertise in Teaching and Evidence Sources Used in Bond et al. (2000)

Dimensions	Effect size	Observation coding	Teacher interview	Student interviews	Questions	Assignment log	Questionnaires
Use of knowledge	.69	√	√				
Deep representations	.89		√				
Problem solving	.77	√	√				
Improvisation	.78		√	√	√		
Challenge of objectives	1.13						
Classroom climate	.6	√					√
Multidimensional perception	.45	√					√
Sensitivity to content	.86	√	√				
Monitoring learning and providing feedback	ns	√	√				
Test hypotheses	.96		√				
Respect for students	.58	√	√		√		√
Passion for teaching and learning	.74	√					
Motivation and self-efficacy	.81						√

For almost all dimensions, Bond et al. (2000) found that NBCTs scored significantly higher than the comparison group. Effect sizes (ES) are presented in Table 2.[4] They found NBCTs to engage in favorable instructional practices to a statistically greater extent, with the exception of monitoring learning and providing feedback. They also found that NBCTs engaged in instructional practice associated with *deep* learning objectives that involve the building of conceptual structures, relations, and abstractions rather than the discrete facts and procedures that characterize *surface*-level objectives. However, despite these observed differences in classroom expectations, differences between groups in the student writing samples were not evident. These writing samples were collected from a nonrandom subset of participating teachers.

Overall, Bond et al. (2000) provided evidence suggesting that the NBPTS-certification process could successfully distinguish teachers who varied in their instructional practices outside of the assessment context. They differed in ways that were consistent with many of the characteristics of teaching that are valued in the NBPTS assessment. However, the evidence cannot be considered conclusive because of the methodological limitations noted.

Three more recent studies have also reported on teacher instructional performance. O'Sullivan et al. (2005) attempted to study assessment practices of NBCTs compared to their peers. They drew from a population in North Carolina across teaching fields and NBPTS certifications. Because they were looking at assessment practice across teaching fields, they first attempted to develop a common set of assessment principles that were embodied by all NBPTS standards documents. While each set of standards articulated common principles, they were often described with different language, organization, and detail. Therefore, O'Sullivan et al. reviewed the standards and designed a common framework that included the following principles (the final bullet is a composite of miscellaneous functions of assessment):

- Assessments reflect valued student learning goals.
- Assessment processes and results serve a variety of appropriate purposes.
- Learning goals are translated into assessments that yield accurate results.
- Assessment information is managed well and interpreted correctly.
- Assessment results are communicated effectively.

- Other functions related to assessment are covered as well (e.g., test prep, technology use, advocacy).

Using this framework, O'Sullivan et al. (2005) developed a sample of 250 NBCTs and sent out an initial survey and request for participation in a more elaborate study. In addition to questions about assessment knowledge and practice, they asked each teacher to identify any colleague teaching in the same school and grade to create a comparison group. Slightly more than half of the NBCTs completed the survey. The identified colleagues were sent the same survey and request for participation. Fifty-two of these individuals responded. The study examined both survey results and samples of assessment practice for 45 teachers each in the NBCT and the non-NBCT groups.

Obviously, with the significant nonresponse of NBCTs and the peer group, the generalizability of the results is problematic. A second limitation is that the nature of the comparison group was not well defined. While the colleagues were in the same schools and grades, there is no way of knowing the extent to which the comparison group could meet NBCT requirements had they participated in the NBPTS process.

The survey included questions about the teacher's confidence in conducting assessment, grading practices, demographic information, and a 20-item test about assessment knowledge. All participants were also interviewed. Survey participants were also asked to submit collected samples of classroom assessments and logs of their assessment practice over a 1-month period, which were then rated on a set of dimensions based on the study's framework. Ratings were made without knowledge of NBCT status.

Few differences were found between the two groups on the survey and interview. Although NBCTs rated their own understanding of assessment higher than their peers, there were no differences in the test of assessment knowledge, nor on other questions about assessment practice.

However, there were consistent differences observed from the submitted assessment samples. The assessments of NBCTs were judged to be clearer and more appropriate in purpose (ES $\approx$ .5)[5] and evidenced greater involvement of students in the assessment process (ES $\approx$ .5). The NBCT assessments were characterized by student responses that were more accurate with respect to content (ES $\approx$1). Finally, there was an indication that the learning goals embedded in

the assessments of NBCTs were of higher quality, though this finding was not statistically significant.

On the basis of the logs, the teachers did not differ in the assessment techniques they used. Both groups were equally likely to use multiple-choice tests, group projects, essays, portfolios, etc. There were also no differences in how the groups communicated the results of the assessment. However, NBCTs were judged to be better at avoiding assessment bias caused by poor test item design and less likely to conflate nonacademic factors such as behavior and attendance with grades (ES ≈ .5).

Smith, Gordon, Colby, and Wang (2005) asked whether the classroom assignments and instruction of NBCTs were designed to produce cognitively deeper student responses. The study included 64 teachers from 17 states who had all participated in the NBPTS assessment. Slightly more than half were certified and the rest had not been successful. Teachers came from one of four certification fields: Middle Childhood/Generalist, Early Adolescence/English Language Arts, Adolescence Young Adulthood/Science, and Adolescence Young Adulthood/ Social Studies-History.

Using the same SOLO taxonomy as used by Bond et al. (2000), the researchers characterized instructional plans and assignments as being designed to elicit *deep* or *surface* levels of understanding. Although most teachers' assignments and instruction (64%) were aimed at surface learning outcomes, there were observed differences between the certified and noncertified groups. Approximately one half of the certified teachers articulated instruction targeted at deeper levels of understanding, compared with approximately one fifth of the teachers who were not successful.

The Smith et al. (2005) study shares many of the methodological strengths and weaknesses with Bond et al. (2000). For both, the comparison between NBCTs and unsuccessful candidates allows for a more direct interpretation of differences associated with certification status. However, the Smith et al. samples of NBCTs and unsuccessful candidates were small and nonrepresentative and the authors did not take into account differences in student characteristics.

The final study reviewed in this section, McColskey et al. (2005) compared three groups of fourth and fifth grade teachers from three school districts in North Carolina. The first group consisted of Middle Childhood/Generalist NBCTs (n = 21). The other two groups were teachers who were identified as being high or low on an index of teacher effectiveness determined by a

value-added methodology used in the state. Each teacher was assigned a teacher achievement index (TAI) based on student performance on the state reading and mathematics achievement test, taking into account prior year performance of the students as well as demographic characteristics.[6] The researchers identified teachers who were at the relative extremes of the TAI distribution to form the high (n = 16) and low (n = 14) non-NBCT groups.

Teachers in the three groups were then compared on a variety of measures including teacher surveys of efficacy; interviews about instructional planning and assessment practices; classroom observations of the cognitive demand of student and teacher questioning, student behavior, and classroom management; and quality of reading comprehension assignments and teacher effectiveness ratings by trained classroom observers.

Relatively few significant differences were observed in the study, which is not surprising given the small sample sizes, and therefore, low experimental statistical power. NBCTs did tend to exhibit better planning, though the difference was not significant. NBCTs did evidence significantly greater cognitive challenge in the reading assignments they assigned (ES ≈ 1).[7] However, the high TAI group exhibited statistically greater observational scores in classroom management, instructional organization, fostering positive relationships, and encouraging responsibility for students (ES ≈ 1). A consistent finding is that NBCTs demonstrated less variability across dimensions than did the low non-NBCT group, attributable to the lack of very low scores on any dimension.

Summary of Studies of Instructional Practice

Four studies are reported that examine the instructional practices of NBPTS teachers compared with their peers. Each of the studies has certain methodological limitations, particularly with respect to sampling representation and accounting for student characteristics in these classrooms. Therefore, it is not possible to claim that NBPTS teachers, in general, would engage in stronger instructional practices regardless of what schools they taught in and with which students.

But there are claims that do seem appropriate to make. Given the classrooms and students of teachers in these studies, there is a pattern of evidence that indicates that NBCTs were more likely than other teachers to give challenging assignments, to expect a deeper level of thinking from their students, to encourage more active responsibility for learning from their students, and to plan their instruction more purposefully.

Studies of Impact on Student Achievement

The NBPTS research receiving the greatest amount of public attention focuses on student achievement. While the most basic question is whether NBCTs produce gains in student learning that are greater than those produced by non-NBCTs, there are actually a number of unique questions subsumed by this general topic. In addition, this set of research studies raises issues of appropriate comparison groups, measures of student learning, models and assumptions for considering teachers' and fellow students' contributions to student learning, and other important methodological issues that are bound to surface in explorations of complex educational interventions such as NBPTS.

The studies included in this section follow a general conceptual methodological framework, though they also vary in important ways. First the section reviews the following commonalities:

- *Assuming detectable teacher effects.* The studies assume that through particular methodologies, particularly those in the broad class of value-added methods (Braun, 2005), causal relationships between teachers and student outcomes can be established. These VAMs continue to be the subject of significant critical examination (Braun, 2005; Lissitz, 2005; McCaffrey, Lockwood, Koretz, & Hamilton, 2003).

- *Looking at change.* As previously discussed, it is not appropriate to attribute teacher effectiveness to the achievement levels of a particular cohort of students. In light of established findings of nonrandom assignment of students to teachers, simple achievement differences only confirm that higher-achieving students are taught by particular teachers. Only by examining increases in student learning while under the auspices of a particular teacher can attributions of teacher effectiveness begin to be made. Thus, these methods all use some form of learning gain metrics to identify teacher contribution.

- *Accounting for contextual effects.* The optimal way to evaluate teacher effects would be to conduct an experiment in which teachers are randomly assigned to schools and classrooms. Of course, teachers are not randomly distributed for a whole host of factors. Assuming that schools have some independent influence on student

outcomes, then teacher impact must be estimated above and beyond the general effect of the school itself.

Assuming that students with certain demographic profiles will be more or less likely to demonstrate academic growth, then individual and group student characteristics must be considered when estimating teacher effectiveness. Therefore, most approaches to estimating teacher effects impose some sort of statistical controls for nonrandom assignment of teachers to students and schools. Some researchers also attempt to control for teacher characteristics independent of NBPTS status in order to estimate NBPTS effects uniquely. The precise nature of these controls varies across studies and certainly affects study outcomes and interpretations.

Clotfelter, Ladd, and Vigdor (2006), using data from North Carolina fifth grade teachers, explored two factors likely to influence the distribution of teachers with respect to students of different academic ability. *Teacher sorting* referred to the choices that teachers made regarding the schools for which they would seek employment. There is a consistent literature (e.g., Ingersoll, 2004) pointing to the fact that more qualified teachers are more likely to seek employment in schools that have students with higher achievement levels. *Teacher shopping* referred to the actions that consumers (i.e., parents) take to ensure that their children have access to the most qualified teachers in the school.

Accounting for these two factors in interpreting NBPTS effectiveness and achievement data is necessary, as demonstrated through the following example. Let us assume that there are two schools, A and B, in which A has dramatically higher levels of student achievement than B. It is also the case that school A has a large number of NBCTs and school B has none. Differences in performance cannot be attributed to NBCTs, because the question cannot be addressed of how students in school B would have done had they had access to these same teachers, nor is it known how students in school A would have performed with the school B faculty. Within schools, a parallel confounding is possible if teacher shopping is prevalent, indicating that student achievement and teacher characteristics are not randomly distributed.

Understanding the distribution of NBCTs raises important methodological issues for any understanding of the relative effectiveness of NBCTs on student achievement and instructional practice. As this paper shows, the findings of studies that fail to address distributional issues are much more limited than those that attempt to control for distribution through design and/or analysis.

The studies in this section ask about three facets of NBPTS effectiveness:

1. *To what extent is NBPTS successful in identifying more effective teachers?* The logic underlying this question is that the NBPTS assessment process should be able to distinguish between teachers who are more or less effective teachers. Those who become NBCTs should have an impact on student learning that is measurably different from the impact of teachers who are not successful on the assessment.
2. *To what extent does the NBPTS process help teachers to become more effective?* This question explores whether participation in the process actually improves teacher effectiveness. The critical comparison here is to examine teacher effectiveness prior and subsequent to participating in the assessment process.
3. *To what extent do NBCTs influence overall teaching effectiveness in their schools?* This question inquires whether, through mentoring or other forms of peer leadership, there are observable effects on student learning in schools with NBCTs.

Ultimately, decisive answers to these three questions can be used to evaluate the cost and benefits of policies associated with the NBPTS program. Policies must not only consider educational benefit, but also costs in absolute and relative terms. Would money be better spent on other kinds of initiatives, for example? This paper does not examine these analyses for several reasons. First, the issue of educational economics, and the analysis of assumptions underlying incremental costs of a program like NBPTS, is far too complicated to consider in any serious way in this review. Second, the estimates of benefits from these studies are sufficiently imprecise to make any resultant cost-benefit model extremely speculative.

This section begins by reviewing several studies that use a variety of methodologies that are not sufficiently powerful to address the issues at hand. Yet, because the studies have received attention, the findings are reviewed her. The section then looks at more statistically sophisticated models that are not without their own set of problems.

The previously reported McColskey et al. (2005) study used a hierarchical regression model to estimate student achievement while taking into account student demographic and school characteristics. They found no achievement differences between NBCTs and other teachers in the study. However, the study included only 25 NBCTs from across three school

districts in North Carolina. The student and school controls in the model were also limited, as many of the variables contained missing data.

In a relatively well-publicized study, Vandevoort, Amrein-Beardsley, and Berliner (2004) used a norm-referenced test, the *Stanford Achievement Tests, Ninth Edition* (SAT-9), to examine gain score differences between NBCTs (Early and Middle Childhood/Generalists) and non-NBCTs in Arizona. The researchers included all NBCTs in grades 3–6 for four different annual cohorts. They developed adjusted gain scores by covarying the prior year test score from the current year test score in reading, mathematics, and language.

Vandevoort et al. (2004) reported that a large majority of comparisons across grades and years favor NBCTs (overall effect size =.12). However, as with other studies, these findings must be discounted due to a variety of methodological weaknesses. First, the unit of analysis is the student, not the teacher. So, for example, they report that 121 students of NBCTs in grade 3 had a mean adjusted gain of 34.5, while the 15,231 students of non-NBCTs had a mean adjusted gain of 28.5.[8] Students, though, are not independent treatment groups. Particular groups of students are associated with a particular teacher and classroom in these analyses. Because students are nested in classrooms, it is inappropriate to report the data in the way that they do. Additionally, the number of teachers is exceedingly small—only 35 NBCTs across all grades participated. As an example, assuming there are 20 students per teacher, the finding above references the impact of 6 NBCTs. The study also fails to consider the characteristics of schools and classrooms in which NBCTs and non-NBCTs are located. Additionally, using a single year covariate is likely to lead to a very poor estimate of student achievement gain (Sanders, 2006).

The previously cited study by Smith et al. (2005) examined two types of student outcome measures. The first took a random sample of six students' work from participating teachers and then evaluated the depth of student responses using the SOLO taxonomy. The researchers found that 78% of all student responses were at a surface level. But the proportion of deeper level responses by students of NBCTs was approximately twice the proportion of those who did not succeed on the assessment. This difference, however, was not statistically significant.

The second measure was based on a pair of responses to two essay prompts, one informative and the other persuasive, sent to participating teachers to administer to their students. Students of NBCTs did significantly better than those of participating nonsuccessful NBCTs. However, the results are problematic in that student responses came from only 18 teachers,

evenly divided into NBCTs and nonsuccessful candidates. Further, as with the previous study, results were reported at the student level, not the appropriate teacher level. Finally, Smith et al.'s (2005) analysis did not control for student and school effects.

The Bond et al. (2000) study on teacher practices also collected student writing samples as an attempt to examine student outcomes. No differences in student achievement were observed. This study also did not statistically control for school and student effects, though attempts were made to match teachers in the two groups on as many characteristics as possible.

Two other studies that have received some attention over the years have such methodological weaknesses that they are simply acknowledged as having been considered during this review. Stone (2002) reported on 16 teachers with no consideration of any contextual factors. Amrein-Beardsley and Berliner (2006) used methods critiqued above in Vandevoort et al. (2004). The failure to report sample sizes in comparisons makes it impossible to reasonably interpret the findings.

A number of studies use production function analysis, a methodology drawn from the field of economics. The production function is a model that indexes student learning as the change in student test scores across years. Student learning is viewed as a function of a set of other variables including characteristics of the other students in a classroom, characteristics of other students in the school, characteristics of the teacher independent of NBPTS status, and finally, NBPTS status. The exact nature of characteristics included in each of the major variables is often determined by the data available to the researchers.

When researchers look at teacher effectiveness, they typically include a set of models, each increasing the included set of variables. The most basic model is to simply look at the relationship between NBPTS status and student learning. However, such an analysis does not consider student, school, and teacher effects. Thus, alternative models are also included to more precisely determine the unique contribution of NBPTS certification to student learning.

Clotfelter, Ladd, and Vigdor (2006), cited above, explored teacher effectiveness using a production function approach. Given their findings that NBCTs were more likely to be teaching students with a history of higher achievement, it is not surprising that they found a significant relationship between being an NBCT and student learning for reading (but not for mathematics scores). However, once teacher sorting and teacher shopping were controlled for by including

school and student effects in their model, the relationship was found to be reduced by 50% (ES for reading = .03–.05).

Clotfelter, Ladd, and Vigdor (2006) also showed that controlling for student demographic variables may be insufficient to understand completely the relationship between teaching and student learning. When the authors included information on the extent to which students engaged in activities such as reading (positively associated with learning) and television watching (negatively associated), teacher effects were reduced even further.

There are several notable features of this study. First, the comparison group is problematic because the population of non-NBCTs includes, presumably, many teachers who could have become NBCTs had they pursued certification, or even teachers who will become NBCTs at a later date. Thus, the comparison groups do not cleanly allow for an analysis of the differential effectiveness of teachers who meet and do not meet NBPTS standards.

Second, the number of NBCTs is very small. Clotfelter, Ladd, and Vigdor (2006) used a sample of over 3,000 teachers. Yet, because they report that only slightly more than 3% were NBCTs, the number of NBCTs upon which this analysis is based seems to be about 108.

With likely heterogeneity in comparison group makeup, and such a small number of NBCTs, it would be surprising if large teacher effects were observed. And indeed, the effects reported are either not observed, or small and positive when observed. Under certain model assumptions, the effects reach statistical significance, yet under others, they do not.

The final point to make about this and other studies is that the research captures only a portion of the impact of NBPTS certification. In this case, the data are limited to fifth grade teachers in North Carolina. This means that only one certification, Middle Childhood/Generalist, is relevant to the analysis. Thus, any claims about NBPTS effectiveness are further limited to one certification area in one state and to a certification area that targets generalist teachers, not those with certifications in specific disciplines.

Goldhaber and Anthony (2007), using a production function methodology, also looked at North Carolina data of teachers in grades 3–5. They were able to look at three cohorts from 1997 through 1999. These were years that spanned very significant growth of NBPTS, such that the numbers of NBCTs included in the analysis were 11, 77, and 215 aggregated across all three grades for the years 1997, 1998, and 1999, respectively.

Three years of data allowed Goldhaber and Anthony (2007) to address a series of questions. Teachers were classified as having a particular NBPTS status for each year of data collection. Status first involved whether or not an individual was an NBPTS applicant. For a given year, a teacher could be a future applicant (e.g., 1997 data for a teacher who applied in 1999), a current applicant (1998 data for a teacher who applied in 1998), a past applicant (1999 data for a teacher who applied in 1998), or a nonapplicant (the majority of teachers). NBCT status was assigned similarly—as future, current, past, or non-NBCT.

Classifications of NBPTS status were used to explore a large set of questions, some more convincingly than others, due to the size of the sample involved in the respective comparisons.

To what extent do students of teachers who become NBCTs at some point demonstrate greater achievement gains than those who do not? Here, Goldhaber and Anthony (2007) examined student test gains by looking at NBCT effectiveness prior to the year in which they attained certification. In both mathematics and reading, under all model conditions, significant, but modest achievement gains were attained by students of NBCTs (ES ≈ .05). This was true even when other teacher attributes, such as their teacher test scores were controlled.

The primary analysis compared NBCTs to the rest of the teacher population, with the consequent comparison group ambiguity described in reference to prior studies. However, Goldhaber and Anthony (2007) were also able to compare NBCTs with those who applied and were not successful. Despite the small sample sizes of candidates who did not receive certification (approximately one half the sample size of successful candidates), effects were still observed and were quite a bit larger for mathematics (ES ≈ .05 for reading and ≈ .09 for math).

To what extent does involvement in the NBPTS assessment process contribute to improved teacher effectiveness? Goldhaber and Anthony (2007) explored the effectiveness of NBCTs during the year of assessment participation. They found a consistent negative effect, which they attributed to the very significant time demands on teachers preparing for the assessment (ES ≈ -.06 for reading and ≈ -.10 for math). They also compared performance prior to and subsequent to the year of certification. They did not observe any differences, but it is important to remember that the numbers of teachers involved in this comparison was very small.

To what extent are there different effects for different groups of students? Goldhaber and Anthony (2007) found effects for students receiving free or reduced price lunch (FRPLs) to be approximately twice the size as those for non-FRPLs. Effects for White and African-American

students were similar to each other. In all cases, effects for mathematics scores were larger than for reading.

Taken together, the Goldhaber and Anthony (2007) results suggest that students of NBCTs demonstrate more learning than do students of teachers who were not NBCTs. These gains appear to be greatest for students at greater educational risk. However, any differences in teacher effectiveness were attributable to the teachers, not the assessment process itself, for their teacher effectiveness values were strongest prior to attaining certification. Interestingly, participating in the assessment process led to a temporary decrease in teacher effectiveness, ostensibly due to the significant demands on candidates' time.

As with previous studies, the small number of teachers involved in particular comparisons is of concern in making precise estimates of any effects. For example, across the 3 years, almost 400,000 student records were matched to teachers. However, only 9,000 were matched to a teacher who participated in the assessment at any point during the 3 years. About two thirds of those students had a teacher who succeeded in becoming an NBCT.

In yet another North Carolina study, Sanders, Ashton, and Wright (2005) studied gain scores in mathematics and reading in two large school districts. This study has markedly less methodological detail than other value-added studies reported in this section, making it more difficult to interpret the research. For instance, basic information such as certification area and the number of teachers involved are missing.[9] Nevertheless, the study did receive a good deal of public attention in the debate about the efficacy of NBPTS-certification.

Sanders et al. (2005) compared results using four different models. Grades 4–8 were each analyzed separately using a database that included 4 years of data (1999–2000 to 2002–2003). Teachers were classified as being NBCTs, having failed the assessment, being future candidates, and never being in NBPTS. Comparisons are made across these groupings to evaluate the same kinds of claims investigated by other studies.

The four models used in the study varied on two dimensions. Two models used student achievement scores and estimated student learning by using the prior year student achievement score as a covariate. The other two models used gain scores directly. While various studies for measuring learning have used each these two different approaches, there is no apparent basis for preferring covariates to gain scores or vice versa, however. Maris (1998) provided a detailed analysis of the two methodological approaches.

The second dimension concerned whether or not the model was hierarchical. Sanders et al. (2005) correctly asserted that students are nested within teachers or classrooms. That is, differences in learning gains associated with a teacher may be attributable to shared characteristics of students in that particular classroom that are not explicitly addressed in the model. Other studies reported in this section did not take into account this clustering of students. A potential risk in not accounting for this clustering is that measurement error is underestimated, leading to overestimates of statistical significance of any reported effects.

Without the hierarchical models, the study found that NBPTS teachers were associated with larger student gains. But, as noted previously, these findings could be attributable to student factors rather than the teacher. When the hierarchical models are used, Sanders et al. (2005) found few statistically significant effects supporting the effectiveness of NBPTS teachers, though the vast majority of comparisons, particularly for the covariance model, were in the direction that would support NBPTS effectiveness.

The question is what to make of the lack of significance and whether this represents a significant departure from the Goldhaber and Anthony (2007) study, which appears to demonstrate a significant but small positive effect of NBCTs on student test scores. Certainly Sanders et al. (2005) took the strong position that:

> …the variability among teachers with the same NBPTS certification is considerably larger than the differences between teachers of different status. Consequently, a student who is randomly assigned to a National Board Certified teacher is not much more likely to get an "effective" (or an "ineffective" teacher) than a student assigned to a teacher who has never been in the NBPTS process (or one who failed certification, or one who may in the future become certified). (p. 8)

The certainty with which the authors state their conclusions is surprising in light of the fact that Sanders (2006) himself has been a strong critic of the methods he uses in this study as often producing misleading conclusions. Specifically, he criticizes the gain score and covariance models he uses in this study as providing unstable estimates of learning and of estimate variance. These models also do not handle missing data, which as Sanders notes, are unlikely to occur at random.

In the Sanders et al. (2005) study, no rationale was provided for using gain scores and covariance models rather than the Education Value-Added Assessment System (EVAAS;

Sanders, Saxton, & Horn 1997), which has been developed and applied widely. This is puzzling given Sanders' (2006) conclusion:

Clearly, the average gain model and any of the analysis of covariance (ANCOVA) models with one score as the predictor should never be used because of the great likelihood of providing severely misleading information. (p. 9)

Thus, Sanders et al. (2005) appeared to overstate the strength of conclusions based on findings of nonsignificance. Given both the small numbers of teachers involved in most comparisons and the acknowledged instability of the achievement measures used in the study, it is inappropriate to draw any definitive conclusions about NBCT effectiveness.

Cavalluzzo (2004) carried out one of the only student achievement studies examining secondary teachers in a particular discipline. The proportion of NBCTs who had received NBPTS certification in mathematics is not clear, however. The study examined 9th and 10th grade mathematics teachers in the Miami-Dade County Public Schools. Cavalluzzo used a series of economic production function models similar to other studies, albeit with a larger set of student and teacher variables than were available to the other state-wide and cross-district studies reported thus far.

Student achievement increases of NBCTs were compared to those of the larger teacher population as well as to those who failed or withdrew from the certification process. For all comparisons under the range of models explored, including those with school and student fixed effects, there was a significant, but relatively small advantage for students of NBCTs (ES ≈ .10). This effect size is somewhat larger than those observed with other studies. In addition, Black and Hispanic students benefited from NBCTs more than did other students. Importantly, as with most studies of NBPTS certification, the sample sizes of NBCTs and of those who did not succeed on the assessment were very small.[10]

Reviewing the student achievement studies presented thus far, some general statements can be made. Once statistical controls are made for the schools in which teachers teach, or for the characteristics of students they teach, there tends to be fairly consistent evidence that students of NBCTs show a relatively modest learning advantage over other comparison groups. These differences appear to be associated with NPBTS successfully identifying teachers that are already effective prior to certification. There is currently no empirical evidence from VAM studies that the assessment process itself makes teachers more effective in raising their students'

achievement. However, there are concerns about the stability of any estimates given the small numbers of NBCTs and nonsuccessful candidates available for these analyses, the appropriateness of comparing successful candidates to nonapplicant teachers, and the limited student achievement data available. All of these factors contribute to imprecise findings about differences that may exist.

The Harris and Sass Studies: A Profound Challenge to Value-Added Methods

All of the preceding research, however, is predicated on the assumption that the respective statewide student accountability measures, as currently implemented, are the appropriate metric by which to evaluate teacher effectiveness. The desire to focus on student achievement is captured by Hanushek and Raymond (2002), "If we are interested in student achievement—as we should be—we simply have to focus on student achievement" (p. 47).While there certainly have been criticisms of the nature of these assessments (Pellegrino, Chudowsky, & Glaser, 2001), the accepted wisdom has been that despite imperfect tests, the use of these outcome measures to gauge teacher effectiveness is appropriate and certainly preferable to using measures of teacher input only (Hanushek & Raymond, 2002). These approaches to measuring teacher effectiveness, of course, are not limited to NBPTS studies.

The critical question is whether these measures are valid, and consistent across states, so that tone can reasonably attribute achievement gains to real gains in student learning—that we are, in some abstract way, comparing apples to apples. If, for example, a different measure of student achievement were used, would the story line about teacher effectiveness change as well? A critical set of recent studies by Harris and Sass (2007a; 2007b) suggests that the answer may be more complex than previously assumed.

Harris and Sass (2007a; 2007b) used a methodology similar to Goldhaber and Anthony (2007). However, the study was conducted in Florida over a 5-year time frame and examined teachers of students in elementary, middle, and high schools. Therefore, the Harris and Sass data included substantially higher numbers of NBCTs than previous studies (e.g., more than1,500 and 1,700 NBCTs used in the analysis of mathematics and reading scores, respectively). As with Cavalluzzo (2004), the proportion of particular NBPTS certifications held by teachers was not reported. Additionally, because Florida enacted a mentor payment system for NBCTs, there was also the opportunity to examine the relationship between student achievement patterns to schools with and without NBCT mentors.

The research addressed four primary questions: are NBCTs more effective; does the NBPTS process influence teacher effectiveness; do NBCTs affect performance of fellow teachers; and does certification provide additional information about effectiveness beyond other available teacher quality indicators?

Florida is an unusual state in that students take two standardized achievement tests annually as part of the Florida Comprehensive Assessment Test (FCAT). One battery of tests is the SAT-9, a national standardized norm-referenced test produced by Harcourt Assessment, Inc. In Florida, the test is referred to as the FCAT-NRT. The second battery, known as the FCAT-SSS, is a criterion-referenced test based on the Sunshine State Standards. Thus, the FCAT-SSS is directly linked to state standards and is used for NCLB accountability reporting.

The first analysis that Harris and Sass (2007a; 2007b) undertook used the FCAT-NRT. They opted for the norm-referenced test because they felt it would "minimize potential biases associated with 'teaching to the test'" (p. 13). Of course, a counterargument could be made that, in a standards-based system, the goal is for teachers to teach to the test and thus, a standards-based test might be a more sensitive gauge of teacher impact. Harris and Sass reported a series of very interesting analyses that vary in terms of questions pursued, grade levels, certification cohorts, and model assumptions. In this review, only a subset is presented.

Looking across grades 3–10 and controlling for school and student fixed effects, they found no overall effect for certification in either mathematics or reading. When disaggregated by grade levels, there was a modest significant and positive effect in mathematics for middle and high school (ES ≈ .08), but no discernible effects in reading. This analysis also supports prior research that teacher effectiveness is reduced during the year of application. Harris and Sass (2007a; 2007b), however, also observed that the reduction in performance may persist. That is, the advantage of having an NBCT is realized before they are actually certified. Potentially troubling from the NBPTS perspective, in some cases NBCTs were actually less effective than other teachers in the state post-certification, not just during the application year. However, these patterns are quite inconsistent across grade levels (elementary, middle, and high school) and for mathematics and reading.

Similar trends, though based predominantly on nonsignificant results, are observed when looking at groups of students disaggregated by race/ethnicity, SES (FRPL), and achievement

level (bottom or top fifth). The only significant finding is that there is a modest achievement advantage for Black students taught by NBCTs prior to their receiving certification (ES ≈ .08).

In looking at the effect of NBPTS on non-NBCTs in the same school, Harris and Sass (2007a; 2007b) analyzed the effectiveness of teachers in a school taking into account either the number of NBCTs or the number of NBCT mentor teachers in the school. A small positive and significant "spillover" effect was noted in mathematics and reading for number of teachers mentoring (ES ≈ .02). However, there was a small but significant negative effect for mathematics when the number of NBCTs, not mentors, was used as the independent measure (ES ≈ -.01).

Taken together, this study supported some previous findings of modest positive impact of NBCTs due to teacher effectiveness differences that existed prior to certification. However, this research was also more negative (albeit inconsistent) about NBCTs in that it also appeared to show that the certification process actually resulted in reduced effectiveness subsequent to the year of application. The authors speculated about possible reasons for the patterns of findings, but they do not offer conclusions.

Harris and Sass (2007a; 2007b) then conducted the same series of analyses with the same students and teachers, but this time using the FCAT-SSS test battery as the outcome measure. The differences in results are striking. Table 3 presents valences, effect sizes, and significance levels for comparisons made using the FCAT-NRT and FCAT-SSS assessments. Positive valences reflect findings in favor of NBCTs. Significance reporting follows Harris and Sass' use of .10 (*), .05 (**), and .01 (***) levels.

The overall NBCT effect for reading across grades 3–10 was positive and significant for the FCAT-SSS. For reading, the effect was significant for both precertification and postcertification. For mathematics, there was a marginally significant positive effect of precertification NBCTs only. Grade-level results for the FCAT-SSS are notably discrepant from those for the FCAT-NR. The finding of lower-achievement postcertification is now mostly positive. For elementary teachers, effects remain nonsignificant or marginal. But, in middle school, one now sees significant positive effects prior to NBPTS-certification in reading. In high school, postcertification effects are now positive for the FCAT-SSS, significantly so for mathematics. Of the 18 contrasts in this analysis, 7 switched valences when a curriculum-linked achievement test was used as the outcome measure. Three comparisons changed from a finding

of nonsignificance to a significant effect and one moved from significantly negative to significantly positive.

When results are disaggregated by race/ethnicity, SES and achievement levels, different interpretations are again apparent. Of the 30 comparisons for each test, one half have different valences. Significant and quite large precertification effects are observed in reading for Black students (ES ≈ .33) and students who receive FRPL (ES ≈ .41).

Table 3

Contrasting Findings From Harris and Sass (2007a; 2007b)

Comparison	Elementary grades 4–5	Middle school grades 6–8	High school Grades 9–10
	Math		
NBCT x preapplication period			
NRT	+*ns*	+.08 **	+.08 **
SSS	*-.19 **	+.23 ***	+ ns
NBCT x application year			
NRT	-ns	-ns	*-ns*
SSS	-ns	-ns	*+.13* **
NBCT x postcertification year			
NRT	+ns	-.13 ***	*-.08* ***
SSS	+.15 **	-.13 *	*+.12* ***
	Reading		
NBCT x preapplication period			
NRT	*-ns*	*-ns*	+ns
SSS	+*ns*	*+.30* ***	+ns
NBCT x application year			
NRT	-ns	-ns	-.06 *
SSS	-ns	-ns	-.25 **
NBCT x postcertification year			
NRT	*-ns*	+.03 *	*-ns*
SSS	+*ns*	+.13 *	+*ns*

Note. Significance reporting follows Harris and Sass' notation. Findings reported in italics represent cases in which the valence of the comparison is different between the two studies.
* $p < .10$. ** $p < .05$. *** $p < .01$.

The story on NBCT spillover effects also changes markedly when the FCAT-SSS is analyzed. While there continues to be a small negative effect on non-NBCTs in mathematics as the number of NBCTs in the school increases (ES ≈ -.09), the observed positive effect of mentors on the FCAT-NRT is no longer observed.

What can be made of the dramatically different results and interpretations observed for the two outcome measures? The natural inclination is to ask which set of scores is more correct. There is probably a better argument to be made for using the FCAT-SSS, as that is the measure that more closely reflects teaching objectives defined by the state. Thus, it is reasonable to expect greater teacher influence on a standards-based assessment such as the FCAT-SSS. Additionally, the design of norm-referenced tests is such that it often captures knowledge and skill most independent of specific teacher effects (Schmidt, Houang, & McKnight, 2005). For tests like the FCAT-NRT, which uses vertical scales that allow for an index of gain across grades on a common scale, the items that allow for such linking are those that are likely most resistant to the effects of a particular teacher. So, a reasonable argument could be made, as Harris and Sass (2007a; 2007b) ultimately concluded in their second paper, that the more NBCT-favorable results observed with the FCAT-SSS are more reflective of teacher impact.

Schmidt et al. (2005) argued that the vertical scale models are predicated on assumptions of content homogeneity—that the "same content underlies multiple measurements and that it is thus legitimate to represent them on the same scale in quantifying outcomes" (pp. 146–147). However, when they actually examined the content of mathematics curricula across elementary grades, it was apparent that entirely new content areas were introduced from time to time, challenging the interpretability of year-to-year score changes as a measure of value added.

Perhaps the most important lesson from the Harris and Sass (2007a; 2007b) studies, and providing empirical backing to the Schmidt et al. (2005) argument, is just how sensitive these VAMs appear to be to the characteristics of the outcome measures. The Harris and Sass results make clear that a test is not a test, and these outcome measures are not replaceable units. There remains much to learn about these models and how they interact with particular outcome measures, an issue to which the concluding section returns. What should be clear though, for the purposes of research, is that despite some widely publicized claims for and against NBPTS, research findings should be treated quite tentatively.

Summary of Studies of Impact on Student Achievement

Given the existing methods and available data, several consistent patterns emerge that relate student achievement to NBPTS status. First, the apparent large achievement differences found in some studies are largely attributable to the nonrandom assignment of teachers and students. NBCTs tend to have students who have stronger historical records of achievement. When assignment factors are controlled, students of NBCTs still tend to learn somewhat more, but the findings are quite modest in magnitude.

Perhaps the most important finding, however, is just how difficult it is to carry out effective studies with readily interpretable results. There are no strong methodological studies that use any outcome measures other than student standardized achievement scores. The use of these measures is driven as much by convenience as by any factor of intellectual merit. Were other indicators of student learning also available on a wide scale, a more robust triangulation of teacher effects would be possible. The stronger methodological studies that rely on achievement scores and value-added methodologies may be far more fragile than previously considered, as evidenced by the instability of findings by Harris and Sass (2007a; 2007b).

Studies of Impact of Professional Development Support Efforts

Several studies in the section above attempted to isolate student achievement effects of participating in the NBPTS process. This section focuses on studies that tend to be more qualitative in nature and that focus on teacher perceptions and knowledge development. Several of these studies primarily have used surveys and related methods to gauge perceptions of participants in NBPTS support efforts. The studies, as a group, are lacking in methodological detail and therefore, are very difficult to interpret.

Anderson, Hancock, and Jaus (2001) conducted a local evaluation of a foundation-supported effort in the Charlotte-Mecklenburg (NC) school system to increase the number of NBCTs. The project, called the Charlotte Collaborative, was a partnership involving the school district, NBPTS, and several universities. The effort included professional development programs, outreach, and recruitment targeted at teachers and administrators; outreach and awareness for the general public; and policy advocacy supporting NBPTS. The evaluation included surveys and interviews of teachers, faculty members, school administrators, project staff, and NBPTS staff. Additionally, focus groups, composed only of teachers, were also

conducted. Finally, the evaluators also collected and reviewed relevant documents from each of the local programs, as well as other relevant documents from NBPTS.

The report provided a synthesis of the evaluators' observations, which were generally extremely positive. They attributed the large number of NBCTs in the district to the success of the program. Though this may well have been the case, it is impossible to make any independent judgments of the results because primary data are not reported in any systematic way. However, based on the review presented by the evaluators, it is fair to assume that individuals who were involved with the project viewed it very favorably.

Rinne (2002) sought to understand the factors that determined how individuals came to apply for certification, how they perceived the application process, and how they reacted to support efforts. The study included a survey and a focus group discussion of NBCTs in Indiana. Unfortunately, because only NBCTs were queried, only characteristics and retrospective perceptions of successful candidates can be understood. No information is available concerning why individuals might not pursue certification and drop out from the process or how support efforts were perceived by those who were not successful. A very similar study, with the same set of limitations, was carried out in Montana (Barfield & McEnany, 2003).

Another set of studies focused on support efforts specifically targeted at helping candidates write the extensive commentaries required by the application. Burroughs (Burroughs, 2001; Burroughs, Schwartz, & Hendricks-Lee, 2000) highlighted the unique challenges to teachers posed by the NPBTS assessment process by studying four teachers engaged in the process of preparing for the assessment. Specifically, NBPTS candidates are asked to engage in a kind of discourse about teaching, through their writing, that is not part of their customary practice. Burroughs argued that:

1. NBPTS constitutes a discourse community
2. Such discourse presents difficulties for many practitioners who view themselves as part of a local community of learners (Burroughs et al. 2001, p. 345).

Burroughs (2001) posited that teachers are part of a community of practice in which most discourse occurs orally and in which a great deal of teacher knowledge is tacit. These characterizations are quite different from expectations embodied in the NBPTS process. To become an NBCT, teachers must write very explicitly about their practice, offering insight into their decision-making throughout the teaching process. Burroughs goes on to claim that the

rhetorical skill involved in descriptive, analytic, and reflective writing about teaching is a "problematic, unarticulated standard of the Board" (Burroughs, 2001, p. 223).

NBPTS writing demands have been offered as one of the significant factors in the consistently low passing rates of African-American candidates. Howard, Ifekwunigwe, Williams, Ullicci, and Webber (2006) hypothesized a potential disconnect between the writing demands of NBPTS and the working knowledge and oral traditions of at least some African-American candidates. Howard et al. were also concerned about other factors that might lead African-American candidates to feel alienated and/or threatened by NBPTS.

To address this, Howard et al. (2006) developed a support program in Los Angeles that included a comprehensive recruitment and support strategy. Recruitment involved a variety of outreach strategies and partnerships to identify and encourage the participation of African-American candidates. Strategies included developing a supportive learning community committed to the success of candidates through the use of mentors, practice, and preparation. One of the key components of the entire support effort was an intensive focus on writing about teaching. Over a 2-year period, the project experienced notable success. The project recruited 20 volunteers each year for a total of 40, of which 36 attended an initial 4-day summer workshop. Eighteen eventually submitted portfolios and eight were successful. The 44% passing rate for African-American candidates (8/18), albeit with relatively small numbers, is equivalent to the first-time passing rate of White candidates in the system.

This review identified only one study that attempted to understand, in any depth, the kinds of understandings that NBCT candidates gained from their involvement in the NBPTS certification effort. Lustick and Sykes (2006) used a quasi-experimental approach to examine what Adolescence and Young Adulthood/Science candidates learn from seeking certification. The researchers asked candidates to examine five NBPTS portfolio-like entries and make judgments about the extent to which performance on the entry provided evidence of meeting particular standards. Judgments were elicited through a 40- to 90-minute telephone interview after the participants were sent a package of portfolio-like responses that would be the basis for the interview. The interviews were then coded by highly trained assessors. The authors used a recurrent institutional cycle research design (Campbell & Stanley, 1966) that allowed for pre- and post-assessment measures for candidates through both longitudinal and cross-sectional comparisons.

The authors identified a number of methodological shortcomings, including the self-selected sampling of candidates. Also, trained assessors exhibited low inter-rater agreement of judgments about the 13 specific standards (mean reliability = .46, range .26–.63). Another potential problem is that candidates were asked to review and make comprehensive judgments on a significant number of dimensions during the relatively short interview.

Acknowledging these limitations, some interesting results were identified. In general, they found reasonable support that NBPTS participation enhances a teacher's ability to make standards-based judgments of practice (ES = .47), particularly with respect to two sets of standards: *Advancing Student Learning in Science* (ES=.48; [Science inquiry; Goals and conceptual understanding; Contexts of science]); and *Establishing Favorable Contexts for Learning* (ES = .52; [Engagement; Equitable participation; Learning environment]). Marginally significant differences were found in two other sets of standards: *Preparing the Way for Productive Student Learning* (ES = .34; [Understanding students; Knowledge of science Instructional resources]); and *Supporting Teaching and Learning* (ES = .44; [Family and community outreach; Assessment; Reflection; Collegiality and leadership]).

Thus, this study provides some preliminary evidence that candidates are learning some of the practices of the NBPTS community, as described by Burroughs (Burroughs, 2001; Burroughs et al., 2000). It is important to recognize, of course, that any learning gains observed in this study are specific to a teacher's ability to analyze teaching practice. No information is available as to whether the certification process actually has an influence on the teacher's performance in the classroom.

Freund, Russell, and Kavulic (2005) studied aspects of the mentoring process as candidates pursued certification. They attempted to address three issues: characteristics of mentors and the mentoring process, the relationship of mentoring to NBCT success, and the effects on candidates of particular mentoring communities. The study consisted of three phases and was limited to Early Childhood and Middle Childhood Generalist and Exceptional Needs certification areas.

The first phase began with the recruitment of teachers. From a stratified, random sample of 4,000 candidates, 816 agreed to participate. Candidates were then asked to supply contact information of mentors. One hundred and fifty-nine mentors agreed to participate in the study. Thus, a limitation of the study was that the sample was no longer representative of the population

and no information was reported regarding the comparability of the study sample to the much larger initial sample that received recruitment information.

Phase II consisted of a series of surveys that included a survey of candidates that asked about themselves, their schools, and their mentoring support; a survey of mentors that focused on personal characteristics as well as their mentoring background; a teacher efficacy scale (Gibson & Dembo, 1984) to gauge candidates' and mentors' perceptions of their own teacher efficacy; an epistemological questionnaire (Schommer, 1998) to sample candidates' and mentors' beliefs about the nature of knowledge and learning; an observational system of submitted videotapes for a subset of candidates for evaluating teacher-exhibited warmth and classroom control (Borich, 2003); and a mentor quality scale designed by the research team.

The research provides a broad array of descriptive information about the characteristics of candidates and mentors in this study. While the study found no relationship between mentoring characteristics and candidate achievement, it did find that those who were successful candidates consistently rated their mentors and mentoring experience significantly higher (ES ≈ .30). Successful candidates more regularly engaged with their mentor in activities that were directly linked to the demands of the NBPTS assessment process. Successful candidates also had higher teacher efficacy scores than less successful candidates (ES ≈ .25). No statistical differences were observed in epistemological beliefs and classroom warmth and control.

Freund et al. (2005) included Phase III, which was actually an independent study unrelated to the first two phases. Through interviews of teachers, mentors, and administrators, an attempt was made to understand, retrospectively, the features, strengths, and weaknesses of mentoring systems in Maryland and South Carolina. The interviewees identified factors such as commitment and focus as playing an important role in the effectiveness of professional mentoring communities.

Cohen and Rice (2005) reviewed eight well-established candidate support programs from around the country in order to understand the nature and impact of the programs and their associated costs. Program directors were surveyed and interviewed about characteristics of their respective programs, including sponsorship, candidate population characteristics, mentor characteristics, mentor-candidate matching, information and recruitment, the nature of large and small group mentoring meetings, other mentoring activities, precandidacy programs, specialized assistance to advanced candidates, research and development-related activities of the program,

state and district financial incentives, and related professional development activities. For each of these dimensions, program variation was extremely broad. In some cases, one program would invest significant time and resources in a particular activity, while another program would not attend to the activity at all. They did find that the support models were based on features and principles closely aligned with practices that have been identified in the literature as being consistently important to effective teacher professional development.

Cohen and Rice (2005) then attempted to use this comprehensive description of programs to relate particular features and NBCT outcomes. This aspect of the research was less successful as support program characteristics were confounded with the programs themselves and most particularly, the candidates who entered into the program. Often, comparisons of interest, such as the performance of minority candidates, were based on extremely small sample sizes. The design did not allow the researchers to isolate independent effects of any of these program features either, features that were often highly correlated. With these caveats in mind, the authors found that the candidates most successful in achieving certification were in support programs that had more intense group sessions, had candidates meet with multiple mentors as needed, linked candidates and mentors by certificate area, and required mentor training.

The authors then tried to quantify what the true costs were of these different support programs by examining more closely four sites. They separated costs into three major categories. *Program-related components* consisted of costs to administer the program, including administration, recruitment, meetings, etc. *Process-related components* consisted of costs associated with the actual preparation of the assessment materials, including the candidate fee. *Other components* consisted of mentor training and research and development costs.

The researchers also partitioned costs into compensated and uncompensated costs. Thus, the time a teacher took to complete a portfolio was almost always uncompensated, but calculated as the product of hourly compensation and the average number of hours to complete. All costs were reported on a per candidate basis.[11]

The four districts examined varied in the number of candidates they supported. The San Antonio program supported approximately 9 candidates annually, while the other three programs (Mississippi Gulf Coast, Stanford, and Winston-Salem) were much larger, supporting between 60 and 100 candidates per year. The San Antonio program had a significantly higher total cost ($31,000/year), presumably because of the higher fixed program-related costs distributed over a

much smaller number of candidates. The other three programs' total costs ranged from $18,000 to $23,000. However, once uncompensated costs are factored out, these three support programs had relatively consistent costs between $3,000 and $4,000 per candidate. The San Antonio program, excluding uncompensated costs, was significantly higher (approximately $12,500 per candidate). Programs also differed in how costs were covered, with contributions primarily coming from the respective states and districts, albeit in different proportions.

The fact that the researchers carefully partitioned costs is important, because it is not immediately clear how to treat uncompensated costs. By far, the largest proportion of these uncompensated costs accrues to the candidate. If this is something the candidate desires to undertake, and there is no opportunity cost regarding other rewards the candidate foregoes, should this be viewed as a true economic cost? Ignoring uncompensated costs, it appears that a relatively comprehensive support program can be developed for between $3,000 and $4,000 per candidate. This estimate assumes a relatively sizable number of candidates that allows the program to achieve economy of scale with regard to program-related costs.

Summary of Studies of Impact of Professional Development Support Efforts

Taken together, the studies of NBPTS support efforts provide useful descriptions of the characteristics and range of program implementations. Many support efforts are thoughtfully designed, with specific attention to the kinds of understandings and skills that are important for successful candidacy. There is emerging evidence that some of these programs are important to candidates' success. However, there are sufficient methodological gaps in each of the reviewed research designs that strong conclusions are difficult to make about exactly how candidates benefit from these programs. Nevertheless, the studies are sufficiently well documented that they provide a basis for pursuing more rigorous exploration of the important questions identified.

Studies of Impact of NBCTs Beyond the Classroom

There is very limited research on the extent to which there is NBPTS impact beyond the classroom. Several studies that examined student achievement attempted to isolate "spillover" effects to non-NBPTS peers have been reviewed and critiqued earlier (e.g., Harris & Sass, 2007a; 2007b). The studies reported in this section focus on the leadership roles that NBCTs may take as a result of their professional accomplishment.

Belden (2002) surveyed the population of California NBCTs about the entire NBPTS experience, including their perceptions of the assessment and its impact on their teaching and career. Although most NBCTs felt that NBPTS helped them become more effective in the classroom, few felt it contributed to their ability to work with parents and community. NBCTs also felt that becoming certified did not open up new leadership opportunities. Rather, becoming an NBCT affirmed their leadership roles in which they had been engaged prior to certification.

Sykes et al. (2006) focused on a very specific aspect of leadership, the extent to which NBCTs helped other teachers in their schools more than non-NBCTs. They accomplished this by first surveying a random sample of NBCTs in Ohio and South Carolina and then surveying the full set of teachers in 47 schools across two states. Schools selected for this more intensive survey varied in the number of NBCTs in the school. They used a sophisticated methodology to estimate the effect of certification itself, controlling for differences that might already exist between NBCTs (n = 175) and non-NBCTs (n = 1,095).

Sykes et al. (2006) found quite strong effects (ES = .48), amounting to NBCTs helping approximately 0.6 more teachers than non-NBCTs. This represented a large statistical effect and, as the authors argued, if there were multiple NBCTs in a school, the impact could be quite significant. They speculated that the effects could be due to the NBPTS process itself, which encourages collaboration, often involves professional development support activities, and provides public recognition of certified teachers. An important caveat is that the precise nature and effectiveness of the reported helping behavior is not elicited from the data collection.

Koppich et al. (2007) conducted case studies of nine low-performing schools in California, North Carolina, and Ohio in order to better understand the extent to which NBCTs played leadership roles in their schools and to identify the factors that influenced the roles they played. Interestingly, the researchers found that in most schools, NBCTs did not exercise a significant leadership presence, attributing it to two factors: (a) the reluctance of most principals to encourage leadership through expanded assignments; and (b) NBCTs and non-NBCTs alike adhering to cultural norms of equality among all teachers in a school. Public acknowledgement of special expertise or status was seen as a violation of cultural norms in many schools.

Summary of Studies of Impact of NBCTs Beyond the Classroom

The very limited research in this area is mixed. While one study found that NBCTs were more likely to work with and support other teachers, other studies suggest that such differences

are not necessarily a function of becoming an NBCT. NBPTS may be an effective signal of leadership qualities, but the ecology of many schools may work against providing increased leadership roles as a result of attaining certification.

Studies of the Distribution of NBCTs

Certainly, the significant public investment in NBPTS carries with it the societal obligation that students from all segments of the population should benefit from a program designed to identify and promote accomplished teaching. The research examined in this section addresses the extent to which NBCTs are teaching in schools with different student populations and academic profiles. Several studies also explore factors contributing to NBCT distribution.

Clotfelter et al. (Clotfelter et al., 2004, Clotfelter, Ladd, & Vigdor, 2006) found that, indeed, there was evidence of significant teacher sorting and teacher shopping. NBCTs were more likely to be employed in schools with students who had higher achievement levels on state standardized tests, whose parents were more affluent, and whose parents were more likely to be college graduates. Even within schools, NBCTs were more likely to teach higher-achieving students than their non-NBCT peers. This uneven distribution has an impact on the interpretation of teacher effectiveness data.

Goldhaber, Choi, & Cramer (2007) also studied distributional patterns of NBCTs in North Carolina, including all primary grade teachers with self-contained classrooms. Looking at 4years of data, the researchers attempted to also understand whether teachers who became certified were more likely to switch schools either because of self-selection or district recruitment.

Goldhaber et al. (2007) found evidence for the uneven distribution of NBCTs, but the relative magnitudes of these differences are subject to interpretation. For example, NBCTs were found in more affluent and higher spending districts with higher teaching salaries. However, the salary differential between district salaries for NBCTs vs. non-NBCTs was only 5%. There were also differences in the likelihood of teaching minority students. NBCTs have, on average, 1.5% fewer minority students in their school districts than there are in the overall population. However, there is more evidence of "teacher shopping" in that the differences increase to 4.9% at the school level and 6.7% at the classroom level.

Goldhaber et al. (2007) found one overriding factor determining the frequency of NBCTs in a district. Districts that had instituted local financial rewards had larger proportions of NBCTs.

This makes sense, of course, for the presence of such rewards reflects an explicit commitment by the district to supporting the NBPTS program, a commitment that includes district funds as well as staff time to prepare for the certification process.

There were key differences in the achievement levels of students of NBCTs compared to average achievement of students in the state. By definition, students are evenly distributed by quartile when all teachers in the state are included—25% of the students are in the lowest quartile, 25% are in the highest quartile, etc. In contrast, approximately 47% of the students of NBCTs come from the top achievement test quartile. This finding reinforces the earlier idea that achievement differences between students of different teachers cannot be attributed simply to the teacher effectiveness.

All of these findings, as well as those of Clotfelter, Ladd, and Vigdor (2006), must be contextualized by mentioning that, even in North Carolina, the presence of NBCTs is still an uncommon event. Although Goldhaber et al. (2007) claimed that White students are 30% more likely than African-American students to have an NBCT, the likelihood that any student, regardless of color, had an NBCT was about one half of 1% (.52% for White students compared with .40% for African-American students). These estimates were based on data through 2000. With the increase in NBCTs, the present likelihood that students are in classes with NBCTs is greater. In fact, Clotfelter, Ladd, Vigdor, and Wheeler (2006) reported that as of 2004, almost 4% of teachers in high-poverty schools were NBCTs. This represents a significant increase in NBCTs, but is markedly less than the over 9% of NBCTs in low-poverty schools.

Nevertheless, this lack of frequency makes the analysis of teacher movement after becoming an NBCT very difficult to interpret. Goldhaber et al. (2007) found no significant differences in the characteristics of district or school placements prior and subsequent to becoming an NBCT. However, the sample size is exceedingly small. From the entire population studied, only 18 NBCTs moved from one district to another, while only 68 moved between schools.

Thus, even with seemingly very large samples, the ability to draw conclusions about the distribution of NBCTs is tricky. The database used by Goldhaber (2007), for example, included approximately 70,000 teachers, which seemingly would allow for very robust findings. However, in trying to isolate the relatively few NBCTs in the state with particular certifications, even in the

state with the largest number of NBCTs, the number of cases influencing the findings is dramatically reduced.

Humphrey, Koppich, and Hough (2004) and Koppich et al. (2007) employed a different methodological approach to explore the issue of NBCT distribution. Using data from respective state departments as well as information from the National Center for Education Statistics (NCES), they examined distribution patterns in the six states with the highest population of NBCTs. They compared NBCTs with other teachers by looking at the likelihood of teaching in schools with the following three factors:

1. Schools with 75% or more FRPL.
2. Schools with at least 75% minority student enrollment.
3. Schools with low-performance histories as indexed by performing in the bottom 3 deciles of the test for the preceding 2–3 years.

For each of these factors, NBCTs were underrepresented in all states except for California. The more even distribution in California was attributed to the very significant investment in NBPTS certification that the Los Angeles Unified School District has made. What was not examined in this study was the issue of teacher shopping that was highlighted by Clotfelter, Ladd, and Vigdor (2006).

The findings from these three studies suggest that NBCTs are, indeed, teaching in schools that are more affluent, higher achieving, and with a larger proportion of White students. However, these studies also face the problem of unknown comparison groups. Comparisons are made between NBCTs and the larger universe of teachers, many of which are likely to have similar qualifications to those who have gained certification. These studies do not address two related questions:

1. Are NBCTs more equitably distributed in schools than similarly qualified teachers who have not yet become certified?
2. Are NBCTs more equitably distributed than teachers who have sought, but not been successful in achieving certification?

Some part of unequal distribution is likely to be attributable to the relatively low numbers of minority teachers who are NBCTs. Wayne et al. (2004) attempted to better understand factors contributing to participation in the NBPTS process by interviewing and surveying a relatively

equal number of White, African-American, and Hispanic teachers from three states. They included 40 NBCTs, 32 who were not successful, and 14 who were eligible to participate in NBPTS, but did not apply. The researchers asked three sets of questions:

1. What factors enter into teachers' decisions about whether or not to pursue certification and do these factors differ by race/ethnicity of the teacher?
2. How do teachers prepare for the portfolio requirements of the assessment? What aspects of preparation are helpful and what makes the assessment difficult? Do aspects and/or perceptions of preparation vary by race/ethnicity of the teacher?
3. How do teachers prepare for the assessment center requirements of the assessment? What aspects of preparation are helpful and what makes the assessment difficult? Do aspects and/or perceptions of preparation vary by race/ethnicity of the teacher?

The relatively small, uncontrolled sample makes it very difficult to interpret any inter-group findings. The sample did not control or report on any group differences in terms of gender, experience, school level (e.g., elementary), or certification field. Although very few differences were observed, the factors most identified with deciding to become (or not become) a candidate were in order (with % responding affirmatively):

1. Wishing to validate teaching capabilities or increase professional status (49%)
2. Competing home and/or school responsibilities considered (36%)
3. Availability of salary bonus or increase (29%)
4. Availability of assessment fee waivers (21%)
5. Opportunity to improve teaching knowledge and/or skills (15%)

Teachers who had entered into the NBPTS process almost all cited encouragement from administrators, teachers, and/or family as most helpful to preparation for the assessment. Most frequently cited barriers were competing obligations at school and/or home. Findings were consistent across questions focused on the portfolio and assessment center aspects of the candidacy process.

Finally, Berry (Berry & King, 2005; Berry, Ferriter, Banks, & Drew, 2006) brought together large numbers of NBCTs, school administrators, and others (550 in total) from North

Carolina to consider issues and solutions to the inequitable distribution of accomplished teachers, independent of NBCT status. Focus group and listserv conversations yielded suggestions that included providing leadership opportunities to accomplished teachers, providing economic incentives, relaxing bureaucratic demands, increasing autonomy and professional consideration, and improving the general working conditions of high-needs schools.

Summary of Studies of the Distribution of NBCTs

Koppich et al. (2007) quote the initial founders' sense of NBPTS as a system that was likely to illuminate the unequal distribution of accomplished teachers within this country. And, indeed, the research seems to support that prediction. NBCTs are more likely to teach wealthier, higher-achieving students in classrooms with relatively high proportions of White students. There is little evidence that the presence of NBPTS has altered this national trend.

Putting NBPTS Impact Research in Context

The studies included in this review are important to the field of educational policy and evaluation research for a variety of reasons. The present review, and the studies upon which it is based, were all done to address the seemingly straightforward questions of the impact of the NBPTS program in identifying and developing effective teachers. However, combining the complexities of schools and schooling, the NBPTS program, outcome measures, and evaluative methodologies makes clear that a simple, definitive answer about the effectiveness of any intervention as ambitious as that of NBPTS will be elusive.

Arguably, the program of research on NBPTS is the strongest, most coordinated, and most comprehensive body of research about the field of teaching that has ever been undertaken around an independent educational intervention. So, the tentativeness of conclusions reviewed here should not be seen as an abrogation of responsibility to conduct sound research. Indeed, researchers have been extremely diligent in designing studies and securing data that can address these questions. But inevitably, all research efforts have faced important constraints that have limited the power of their conclusions. Thus, the lack of definitiveness about NBPTS impact in this review should not be seen as an attempt to avoid confronting a compelling set of data. Instead, what the research reveals is that the evaluation of systemic interventions in a manner that is sensitive to the complexity of the intervention, and the context in which it occurs, requires

a sustained and evolving program of research. In the current case, an unprecedented foundation has been laid, but it remains only a foundation.

Understanding of NBCT impact continues to be tentative. Though there is evidence that NBCTs are associated with more effective instructional practices, increased student achievement, and stronger leadership, identified effects are not overwhelming. Is this because the NBPTS process does not distinguish stronger from weaker teachers, or might there be legitimate reasons to suggest that the research is still very much in its infancy, requiring continued and refined efforts until a more definitive resolution is reached?

In the following sections, a set of issues are identified that elaborate on the complexities that make the pursuit of these questions both difficult and important. Future research directions are also proposed.

Single Certificate Research Studies as Proxies for the NBPTS System

While almost every reviewed study discusses NBPTS effectiveness as a unitary construct, the fact is that most studies only examine performance with respect to one or two certificates. Additionally, the vast majority of studies, particularly student achievement studies, examine only the generalist certificates. But there are significant differences in the academic background of generalist and content-specific teachers (Gitomer, Latham & Ziomek, 1999). Teachers with content-specific certifications have much stronger academic credentials, on average, than those who seek elementary certifications. Therefore, the current status of research is one in which generalizations are made about NBPTS effectiveness when the only legitimate claim that can be made is with respect to the certificate areas under study.

It may turn out that effectiveness of NBPTS will range across certification areas. Obviously, the choice to study generalists has been governed by the size of available samples. But, as NBCTs become more prevalent, the opportunity to study NBPTS effects across certification areas may help illuminate not only characteristics of NBPTS, but also characteristics of teaching more generally.

The Distribution of Teachers and Limitations of Generalization

The research makes clear that NBCTs are not randomly distributed across all schools and classrooms. Therefore, researchers have rightfully raised the concern that any identified effects may be confounded with distribution patterns, despite statistical controls. There is another

related, but less discussed, feature of the NBPTS assessment that also limits generalizations about the effectiveness of any teacher or group of teachers.

The fact is that the most significant contributor to NBPTS assessment performance is the classroom portfolio that requires evidence from the teacher's actual practice. Thus, teacher quality is defined by the evidence the teacher provides within a single context. Certification is awarded under the assumption that effective teachers have particular qualities that transcend particular contexts. Yet, the assessment and the existing research base have virtually no evidence to evaluate this assumption. Ultimately, a corollary to the question of whether NBCTs work in better schools is whether it is easier to demonstrate qualities valued in the NBPTS certification process with better students. Despite all of the very impressive efforts that NBPTS has developed to assure fairness in judging teaching quality, this fundamental question is not answerable at this time.

Certainly, one could imagine a hypothetical research study in which NBCTs and appropriate controls were randomly assigned to classrooms representing different educational contexts (i.e., school and student characteristics) to address this issue. Whether such a study could be carried out practically, however, is questionable.

The Representation of States

The majority of research reported comes from North Carolina, a state that committed to NBPTS from the beginning of the program. Indeed, Governor James Hunt became the founding Chair of NBPTS during his tenure in office in North Carolina. The recent Harris and Sass (2007a, 2007b) studies are the first large-scale student achievement efforts based on data from another state. Yet, education remains largely a state issue, with a variety of procedures, practices, standards, and traditions. Therefore, evaluation of national programs such as NBPTS will benefit from a broader sampling of data.

Potential Changes in Populations Within States

States have supported NBPTS to differing extents and in different manners. Some, like North Carolina, have made strong, long-term commitments from the inception of NBPTS. Others have made more gradual commitments. Virtually nothing is known about how the teacher candidate pool changes within a state over time. Are there characteristics of candidates who participate early in the process that are different from those who participate after a program has

been in place for several years? How do selection criteria for determining which candidates will be supported change over time? Some of the studies reported made initial attempts to report cohort effects, but little systematic evaluation of the changing nature of candidates has been done.

The Effectiveness of the NBPTS Passing Score

NBPTS and research studies treat certification as a dichotomous classification. However, two facts are relevant in future considerations of both NBPTS policy and research interpretations. First, certification scores of NBCTs vary dramatically. Second, the midpoint of the distribution of candidate scores for most certificates is very close to the passing scores. What this means is that there is a very large proportion of candidates who are very near the passing score, some who get certified and some who do not. Yet, their relative performances are statistically indistinguishable. Any time an assessment program establishes a fixed passing standard, there will be individuals who fall within a range where there is some nontrivial likelihood of classification error. Thus, in comparing NBCTs and those who were not successful, studies may be comparing individuals who have been classified differently, but who performed similarly on the assessment.

As NBPTS has matured as a program, it may be appropriate to reconsider the passing score. Certainly, moving it higher in the score distribution would result in fewer potential classification errors. As the population of NBCTs increases it should also be possible for researchers to draw from candidates who did evidence statistically different performances in the assessment.

The Appropriateness of Comparison Groups

As was noted throughout this review, the bulk of the research relies on comparisons that compare NBCTs to the general population of teachers, an unknown group that undoubtedly includes many teachers who are as accomplished in their teaching as NBCTs, but have not participated in the voluntary certification process. Studies about the impact of NBPTS as professional development would benefit from different comparison groups than, say, research on NBPTS as a signal of effective teaching.

The largest challenge to establishing adequate comparison groups has been the lack of adequate sample sizes of comparison groups with appropriate characteristics. The status of this

situation is likely to change as the number of NBCTs and those who have not been certified increases.

The Adequacy of Outcome Measures

The research studies that have drawn the most attention have been ones focused on student achievement. Indeed, the focus on VAMs of student achievement is becoming the dominant research methodology for evaluating all manner of educational interventions. Yet, the Harris and Sass (2007a; 2007b) studies provide compelling empirical evidence of the risks in treating these measures uncritically.

There is general consensus in the research and policy communities that student learning ought to be the single most important consideration in the evaluation of any educational intervention. The allure of econometric models such as VAMS is inescapable, with student learning gains used as the measure of benefit. But the metaphor with economic models breaks down in some important ways.

First, tests do not yield the same results, even when those tests nominally measure the same thing. Feuer, Holland, Green, Bertenhall, and Hemphill (1999) demonstrated that different achievement tests could not be readily substituted for one another without compromising the validity of interpretations. Changing the measures, within or across states, is likely to change interpretations, and it appears from the Harris and Sass (2007a; 2007b) work, even more significantly than many expected.

Second, the tests used are proxy measures that are not fully representative of learning goals for students as defined not only by NBPTS, but by national standards and models of student learning and cognition (e.g., Pellegrino et al., 2001).

Third, as Schmidt et al. (2005) argued, gain scores implies some year-to-year continuity in terms of what is being measured. In order to have such continuity, test-makers create vertical scales that rely on common items across years. Yet, curriculum varies in very substantial ways across these same years. Thus, the basis for vertical scaling is comprised of tasks that are least likely to be affected by the curriculum-focused teaching activities of a given year (see also Walsh, 2006).

So, the economic metaphor is imperfect in some very critical ways. Imagine a study of the economic impact of educational attainment, where annual income is the outcome measure. Then, contrast this with student achievement gains as the outcome measure. In one case, a dollar

means the same thing in California as it does in North Carolina. A dollar also has common meaning for high school and college graduates. Achievement scores, for reasons noted above, do not mean the same thing for third and eighth graders. They are measuring very different things. With student achievement, there is no common measure. In one case, there are clear external referents—one knows what kind of a car $2,000 will purchase. There is no way of equating that with .2 of a standard deviation of some achievement scale.

The Adequacy of Value-Added Models

Clearly, one threat to the usefulness of these models comes from the utility of these outcome measures and the validity of interpretations that are based on such outcome measures. But, there are many unanswered questions about the VAMs themselves that require much deeper investigation. In the most thorough review of VAMs to date, McCaffrey et al. (2003) concluded

> The research base is currently insufficient to support the use of VAM (value-added models) for high-stakes decisions. We have identified numerous possible sources of error in teacher effects and any attempt to use VAM estimates for high-stakes decisions must be informed by an understanding of these potential errors. (p. 119)

A highly accessible synthesis of VAMs for the nontechnical reader is provided by Braun (2005) and raises similar cautions. VAMs are based on a broad range of assumptions, many relatively unexamined, and others questionable. McCaffrey et al. (2003), Braun (2005), Lissitz (2005), and others who have evaluated these models carefully all agreed that value-added modeling is a promising step towards developing evaluative techniques that focus on student learning. But all also conclude that the methods, and the assumptions upon which they are based, require much more substantial understanding.

Concluding Comments

NBPTS, public and private funding sources, and the researchers cited have committed to a research program that has highlighted two significant challenges—making sense of the impact of NBPTS, but also understanding better how to undertake serious investigation of complex educational interventions. Indeed, significant progress has been made on both fronts. The progress is such though that this paper closes with the conundrum articulated by President John F. Kennedy, “The greater our knowledge increases the more our ignorance unfolds.”

References

Amrein-Beardsley, A., & Berliner, D. C. (2006, April). *The residual effects of National Board certified teachers: An analysis of student achievement trends after students are taught by National Board certified teachers.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Anderson, K. M., Hancock, D., & Jaus, V. (2001, November 19). *Program evaluation report for the Charlotte Collaborative Project involving the National Board for Professional Teaching Standards, Charlotte-Mecklenburg Schools, Johnson C. Smith University, and University of North Carolina at Charlotte* (NBPTS Rep.). Retrieved September 6, 2006, from http://www.nbpts.org/UserFiles/File/Charlotte_Collaborative_D_-_Anderson.pdf

Barfield, S. C., & McEnany, J. (2003). *Montana's national board certified teachers' views of the certification process.* Unpublished manuscript, University of Montana, Billings:.

Belden, N. (2002). *California teachers' perceptions of National Board Certification: Report of findings from a survey for the Center for the Future of Teaching and Learning* (NBPTS and Center for the Future of Teaching and Learning Rep.). Retrieved September 7, 2006, from http://www.cftl.org/documents/Beldenreport2002.pdf

Berry, B., Ferriter, B., Banks, C., & Drew, K. (2006). Every child deserves our best: Recommendations from North Carolina's National Board certified teachers on how to support and staff high needs schools (National Education Association and Center for Teaching Quality Rep.). Retrieved September 6, 2006, from http://www.teachingquality.org/pdfs/ncnbctrecs.pdf

Berry, B., & King, T. (2005). R*ecruiting and retaining National Board certified teachers for hard-to-staff, low-performing schools: Silver bullets or smart solutions* (Southeast Center for Teaching Quality Rep.). Chapel Hill, NC: Southeast Center for Teaching Quality.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy.* New York: Academic Press.

Black, P., & Wiliam, D. (1998). *Assessment and classroom learning. Assessment in Education, 5*(1), 7–74.

Bond, L., Smith, T., Baker, W.K., & Hattie, J. (2000). The certification system of the National Board for Professional Teaching Standards: A construct and consequential validity study

(NBPTS Rep.). Retrieved September 6, 2006, from http://www.nbpts.org/UserFiles/File/validity_1_-_UNC_Greebsboro_D_-_Bond.pdf

Borich, G. D. (2003). *Observational skills for effective teaching.* Columbus, OH: Prentice-Hall/Merill.

Bransford, J. D., Brown, A., & Cocking, R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school.* Washington, DC: National Academy Press.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models* (Policy Information Rep.). Princeton, NJ: ETS.

Burroughs, R. (2001). Composing standards and composing teachers: The problem of National Board certification. *Journal of Teacher Education, 52*(3), 222–232.

Burroughs, R., Schwartz, T. A., & Hendricks-Lee, M. (2000). Communities of practice and discourse communities: Negotiating boundaries in NBPTS certification. *Teachers College Record, 102*(2), 344–374.

Campbell, D., & Stanley, J. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Carnegie Corporation of New York. (1986). *A nation prepared: Teachers for the 21st century.* New York: Author.

Cavalluzzo, L. (2004). *Is National Board Certification an effective signal of teacher quality?* (pp. 1–39). Retrieved May 24, 2007, from http://www.cna.org/documents/CavaluzzoStudy.pdf

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2004). *Teacher quality and minority achievement gaps* (Working Paper Series San04-04). Retrieved September 2006, from the Terry Sanford Institute of Public Policy Web site: http://www.pubpol.duke.edu/research/papers/SAN04-04.pdf

Clotfelter, C. T., Ladd, H., F., & Vigdor, J. (2006). *Teacher-student matching and the assessment of teacher effectiveness* (NBER Working Paper No. 11936). Stanford, CA: National Bureau of Economic Research..

Clotfelter, C. T., Ladd, H. F., Vigdor, J., & Wheeler, J. (2006, December 7). *High poverty schools and the distribution of teachers and principals* (CALDER Working Paper 1). Retrieved April 2007, from the National Center for Analysis of Longitudinal Data in

Education Research Web site: http://www.caldercenter.org/PDF/1001057_High_Poverty.pdf

Cohen, C. E., & Rice, J. K. (2005). *National Board Certification as professional development: Design and cost* (pp. 1–198). Arlington, VA: NBPTS.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenhal, M. W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests.* Washington, DC: National Academy Press.

Freund, M., Russell, V. K., & Kavulic, C. (2005). *A study of the role of mentoring in achieving certification by the National Board for Professional Teaching Standards* (NBPTS Rep.). Arlington, VA: NBPTS.

Gibson, S., &. Dembo, M. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology, 76*(4), 569–582.

Gitomer, D. H., Latham, A. S., & Ziomek, R. (1999). *The academic quality of prospective teachers: The impact of admissions and licensure testing* (ETS Research Rep. No. RR-03-25). Princeton, NJ: ETS.

Goldhaber, D., & Anthony, E. (2007). C*an teacher quality be effectively assessed? The Review of Economics and Statistics, (89)*1, 134–150. Retrieved April 2, 2007, from http://www.mitpressjournals.org/doi/abs/10.1162/rest.89.1.134?cookieSet=1&journalCode=rest

Goldhaber, D., Choi, H., Cramer, L. (2007). A descriptive analysis of the distribution of NBPTS-certified teachers in North Carolina. *Economics of Education Review, 26*(2), 160–172.

Hanushek, E. A., & Raymond, R. E. (2002). I*mproving educational quality: How best to evaluate our schools?* Paper presented at Education in the 21st century: Meeting the challenges of a changing world, Boston, MA.

Harris, D. N., & Sass, T. R. (2007a, January). *The effects of NBPTS-certified teachers on student achievement* (NBPTS Rep.). Retrieved April 2007, from http://www.nbpts.org/UserFiles/File/Harris_Sass_Final_2007.pdf

Harris, D. N., & Sass, T. R. (2007b, March). *The effects of NBPTS-certified teachers on student achievement* (CALDER Working Paper 4). Retrieved April 2007, from the National Center for Analysis of Longitudinal Data in Education Research Web site: http://www.caldercenter.org/PDF/1001060_NBPTS_Certified.pdf

Howard, T., Ifekwunigwe, A., Williams, R. J., Ullucci, K., & Webber, K. (2006, January). *Closing the achievement gap in National Board Certification: Optimal support for African American teacher candidates* (NBPTS Rep.). Retrieved September 6, 2006, from http://www.nbpts.org/UserFiles/File/NBPTS_Final_Report_Adverse_Impact_D-_UCLA_-_Howard.pdf

Humphrey, D. C., Koppich, J., & Hough, H. (2004). *Sharing the wealth: National Board certified teachers and the students who need them most* (NBPTS Rep.). Arlington, VA: NBPTS.

Ingersoll, R. M. (2004). Why some schools have more under-qualified teachers than others. In D. Ravitch (Ed.), *Brookings papers on education policy: 2004* (pp. 45–88). Washington, DC: Brookings Institution Press**.**

Koppich, J. E., Humphrey, D. C., & Hough, H. (2007). Making use of what teachers know and can do: Policy, practice, and National Board Certification. *Education Policy Analysis Archives, 15*(7). Retrieved April 10, 2007, from http://epaa.asu.edu/epaa/v15n7/

Lissitz, R. W. (Ed.). (2005). *Value added models in education: Theory and applications.* Maple Grove, MN: JAM Press.

Lustick, D., & Sykes, G. (2006). N*ational Board Certification as professional development: What are teachers learning? An empirical investigation of the learning outcomes from the National Board for Professional Teaching Standards' certification process* (NBPTS Rep.). Arlington, VA: NBPTS.

Maris, E. (1998). Covariance adjustment versus gain scores—revisited. *Psychological Methods*, *3*, 309–327.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica, CA: The RAND Corporation.

McColskey, W., Stronge, J. H., Ward, T. J., Tucker, P. D., Howard, B., Lewis, K., et al. (2005). *Teacher effectiveness, student achievement, and National Board certified teachers* (NBPTS Rep.). Arlington, VA: NBPTS.

National Board of Professional Teacher Standards. (n.d.). *The standards.* Retrieved May 22, 2007, from http://www.nbpts.org/the_standardsO'Sullivan, R., Hudson, M., Orsini, M., Arter, J., Stiggins, R., & Iovacchini,.L. (2005). Investigating the classroom assessment literacy of NBPTS Board certified teachers (NBPTS Rep.). Arlington, VA: NBPTS.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Redfield, D. L., & Rousseau, E. W. (1981). Meta-analysis of experimental research on teacher questioning behavior. *Review of Educational Research, 51*(2), 237–245.

Rinne, M. G. (2002). *Status of National Board certified teachers in Indiana.* Indianapolis, IN: Indiana Professional Standards Board.

Sanders, W. L. (2006). *Comparisons among various educational assessment value-added models.* Paper presented at the power of two: National value added conference, Columbus, OH.

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

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system, a quantitative, outcomes-based approach to educational measurement. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press.

Schmidt, W. H., Houang, R. T., & McKnight, C. C. (2005). Value-added research: Right idea but wrong solution? In R. W. Lissitz (Ed.), *Value added models in education* (pp. 145–164). Maple Grove, MN: JAM Press.

Schommer, M. (1998). The influence of age and education on epistemological beliefs. British *Journal of Educational Psychology, 68,* 551–562.

Schön, D. (1987). *Educating the reflective practitioner.* San Francisco: Jossey-Bass.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. Harvard *Educational Review, 19*(2), 4–14.

Smith, T. W., Gordon, B., Colby, S. A., & Wang, J. (2005, April). *An examination of the relationship between depth of student learning and National Board Certification status* (NBPTS Rep., pp. 1–193). Arlington, VA: NBPTS.

Sternberg, R. J., & Horvath, J. A. (1995). A prototype of expert teaching. *Educational Researcher, 24*(6), 9–17.

Stone, J. E. (2002). *The value-added gains of NBPTS-certified teachers in Tennessee: A brief report.* Retrieved September 7, 2006, from the Educational Resources Information Center Web site, http://eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/28/02/a8.pdf

Sykes, G., Anagnostopoulos, D., Cannata, M., Chard, L, Frank, K., McCrory, R., et al. (2006). *National Board certified teachers as organizational resource.* Arlington, VA: NBPTS.

Vandevoort, L. G., Amrein-Beardsley, A., & Berliner, D. (2004). National Board certified teachers and their students' achievement. *Education Policy Analysis Archives, 12*(46). Retrieved May 24, 2007, from http://epaa.asu.edu/epaa/v12n46/

Walsh, K. (2006, March). *If wishes were horses: The reality behind teacher quality findings.* Paper presented at the Brookings Institution workshop on measuring child well-being, Washington, DC.

Wayne, A., Chang-Ross, C., Daniels, M., Knowles, K, Mitchell, K., & Price T. (2004). *Exploring differences in minority and majority teachers' decisions about and preparation for NBPTS certification* (NBPTS Rep.). Arlington, VA: NBPTS.

Notes

[1] Financial information compiled and provided by NBPTS, April 2007. In the years 2004 and 2005, NBPTS reported a total income of $84.9 million, of which $18.1 million came from federal sources.

[2] In 2002, NBPTS, with the support of a consortium of foundation funding, issued a request for proposals (RFP) that resulted in the funding of more than 20 projects at a cost of $6.5 million. The administration of the RFP and resulting review process was handled by the RAND Corporation, operating as an independent agent.

[3] The revision also changed the focus of the assessment center. Whereas the earlier version included tasks that measured content and pedagogical-content knowledge, the revision primarily focused on content only.

[4] In many cases, effect sizes (ES) are reported as they were in the original study. However, in cases where multiple models and comparisons are made, an approximate ES is reported that reflects the average of relevant effects observed. In these cases, the approximation is indicated by the ≈ symbol. For precise estimates of all effects, the reader is referred to the original studies.

[5] Effect sizes were not reported in the summary paper cited. They were calculated for the review on the basis of information provided in the cited report.

[6] Issues associated with value-added methodologies will be explored more thoroughly in the section on student achievement.

[7] Effect sizes were not reported in the study. They were calculated for the review on the basis of information provided in the cited report.

[8] The scale is not reported in the research, but a search of the Arizona Web site suggests the development of a state-specific scale of 300-700 for mathematics and reading and 300-900 for language. See http://www.ade.state.az.us/researchpolicy/AIMSResults/scoreranges.pdf.

[9] The study reports teacher-years, which are a product of the number of teachers and the number of years of data available for that teacher in the three-year data file. Across grades and subject areas, the median number of teacher-years was approximately 30 and as low as 8 for certified teachers and 5 for noncertified teachers.

[10] As with Sanders, Ashton, & Wright (2005), the data reported are in terms of teacher-years that included four years of data matched with students. The teacher-years for NBCTs were 61 in total. Teacher-years for current applicants were 101.

[11] There were some relatively small inconsistencies in the reporting of costs that made reconciliation problematic. The differences, however, were not of a magnitude to change the essential points made in the text.

Appendix

Summary of Studies Addressed in This Review

Authors	Publication status	State(s)	Certifications	Comparison groups and size	Research question(s)
Studies of instructional practice					
Bond, Smith, & Hattie (2000)	Technical report	NC	EA/ELA and MC/Gen	65 candidates split among 31 NBCTs and 34 non-NBCTs	Do those who passed and did not pass the assessment differ on 13 dimensions of instructional practice?
O'Sullivan et al. (2005)	Technical report	NC	MC—specialty not specified	52 matched pairs—NBCTs and noncandidate peers	Do NBCTs and other teachers differ in understanding and practices of assessment?
Smith, Gordon, Colby, & Wang (2005)	Technical report	17 states	MC/Gen, EA/ELA, AYA/Science, AYA/SSH	64 candidates split among 35 NBCTs and 29 non-NBCTs	Do students taught by NBCTs produce deeper responses than students of teachers who attempted, but did not achieve certification?

(Table continues)

Table (continued)

Authors	Publication status	State(s)	Certifications	Comparison groups and size	Research question(s)
Clotfelter, Ladd, & Vigdor (2006)	Peer-reviewed journal	NC	MC/Gen	~108 NBCTs (5th grade) compared value-added scores with other teachers	Do NBCTs produce larger learning gains, particularly when taking into account student and school distribution factors?
Goldhaber & Anthony (2007)	Peer-reviewed journal	NC	EC/Gen MC/Gen	~300 NBCTs (Grades 3-5) compared value-added scores with other teachers and those who did not pass	Do NBCTs produce larger learning gains, taking into consideration whether comparisons are made prior to, during, or subsequent to certification and controlling for student and school factors?

(Table continues)

Table (continued)

Authors	Publication status	State(s)	Certifications	Comparison groups and size	Research question(s)
Harris & Sass (2007a; 2007b)	Technical report	FL	Not specified	Large numbers of NBCTs (grades 3–10) (~1,500 for math scores, ~1,700 for reading) compared value-added scores with other teachers and those who did not pass	Do NBCTs produce larger learning gains, taking into consideration whether comparisons are made prior to, during, or subsequent to certification and controlling for student and school factors?
McColskey et al. (2005)	Technical report	NC	MC/Gen	21 NBCTs, 16 with high value-added scores, 14 with low value-added scores	Do NBCTs differ from other teachers in terms of efficacy and instructional practices?

(Table continues)

Table (continued)

Authors	Publication status	State(s)	Certifications	Comparison groups and size	Research question(s)
Sanders, Ashton, & Wright (2005)	Technical report	NC	Not specified	Undefined number of NBCTs (grades 4–8) compared value-added scores with other teachers and those who did not pass	Do NBCTs produce larger learning gains, taking into consideration whether comparisons are made prior to or subsequent to certification and controlling for student and school factors?
Smith, et al. (2005)	Referenced above				Do students taught by NBCTs produce deeper responses than students of teachers who attempted, but did not achieve certification? Are writing samples of students taught by NBCT better than essays from other teachers?

(Table continues)

Table (continued)

Authors	Publication status	State(s)	Certifications	Comparison groups and size	Research question(s)
Stone (2002)	Technical report	TN	Generalist (certification not specified)	16 teachers compared value-added scores with other teachers	Do adjusted gain scores differ for NBCTs and other teachers?
Vandevoort, Amrein-Beardsley, & Berliner (2004)	Online journal	AZ	EC/Gen MC/Gen	35 NBCTs across 4 grade levels—value-added scores compared with general population	Do adjusted gain scores differ for NBCTs and other teachers?
Studies of impact of professional development support efforts					
Anderson, Hancock, & Jaus (2001)	Technical report	NC	Various	282 stakeholders (teachers, faculty, school administrators, mentors) surveys, interviews, focus groups, and document review	What is the nature and perceived impact of an NBPTS professional development support effort?

(Table continues)

Table (continued)

Authors	Publication status	State(s)	Certifications	Comparison groups and size	Research question(s)
Barfield & McEnany (2003)	Technical report	MT	Various	Survey of NBCTs	What are the factors that determine candidate participation, how they perceive the application process and support efforts?
Belden (2002)	Technical report	CA	All	Survey of 519 CA NBCTs (68% of population)	What are perceptions of NPTS process and impact on career?
Burroughs (2001)	Peer-reviewed journal	OH	EC/Gen MC/Gen	Case study of 2 teachers (subset of Burroughs et al., 2000)	What are the discourse practices necessary for NBPTS success and how are they developed?
Burroughs, Schwartz, & Hendricks-Lee (2000)	Online journal	OH	EC/Gen MC/Gen	Case study of 4 teachers	What are the discourse practices necessary for NBPTS success and how are they developed?

(Table continues)

Table (continued)

Authors	Publication status	State(s)	Certifications	Comparison groups and size	Research question(s)
Cohen & Rice (2005)	Technical report	8 sites across states	Various	Interview and examination of 8 sites' professional development efforts, costs, and outcomes	What are the design and cost characteristics of NBPTS support efforts?
Freund, Russell, & Kavulic (2005)	Technical report	National	EC/Gen; MC/Gen; ECYA/ENS	Survey of 816 candidates and 159 of their mentors Interviews of teachers and others involved in support networks	Are characteristics of mentors and mentoring related to candidate success?
Howard et al. (2006)	Technical report	CA	Various	Case study of 40 African-American candidates across 2 cohorts	Can support structures be identified and implemented to increase success rates of NBPTS candidates?

(Table continues)

Table (continued)

Authors	Publication status	State(s)	Certifications	Comparison groups and size	Research question(s)
Koppich, Humphrey, & Hough (2007)	Online journal	CA, FL, MS, NC, OH, SC	All	Survey of 1,136 NBCTs followed by case studies of 9 schools in 3 states	Where do NBCTs teach and what determines their success in influencing educational practice in their schools?
Lustick & Sykes (2006)	Online journal	National	AYA/Science	Structured interviews of assessment understanding of 118 candidates divided among 4 cohorts	Does participating in the NBPTS assessment process improve teacher understanding of student assessment?
Rinne (2002)	Technical report	IN	Various	Survey of 32 NBCTs	What are the factors that determine candidate participation, how they perceive the application process and support efforts?

(Table continues)

Table (continued)

Authors	Publication status	State(s)	Certifications	Comparison groups and size	Research question(s)
			Studies of Impact of NBCTs Beyond the Classroom		
Sykes et al. (2006)	Technical report	OH, SC	EC/Gen; MC/Gen	Survey of 1,153 teachers followed by survey of all faculty in 47 elementary schools varying in NBCT density	Are there discernible effects related to NBCT status on the amount of help given to other teachers in a school?
			Studies of the distribution of NBCTs		
Berry & King (2005) and Berry, Ferriter, Banks, & Drew (2006)	Technical reports	NC	Various	Focus groups of NBCTs, administrators, other teachers, etc.	How can issues of inequitable distribution of teachers be addressed?
Clotfelter et al. (2004)	Peer-reviewed journal	NC			
Clotfelter et al. (2006)	Peer-reviewed journal	NC			

(Table continues)

Table (continued)

Authors	Publication status	State(s)	Certifications	Comparison groups and size	Research question(s)
Clotfelter et al. (2006)	Technical report	NC	All		
Goldhaber, Choi, & Cramer(2007)	Peer-reviewed journal	NC			
Humphrey Koppich, & Hough (2004); Koppich, Humphrey, & Hough (2007)	Online journal Online journal	CA, OH, NC	All	Case studies of 9 schools in 3 states	What factors determine the success with which NBCTs influence practice in their schools?
Wayne et al. (2004)	Technical report	CA, FL, MD	All	Interviews of 86 African-American, Hispanic, and White teachers – successful candidates, unsuccessful candidates, and eligible noncandidates	What factors enter into decision to pursue certification? How do candidates prepare? Are there differences associated with demographic group membership?