2010

Meta-analysis comparing student outcomes for National Board certified teachers and non-National Board certified teachers

Wendy Hacke

The University of San Francisco

META-ANALYSIS COMPARING STUDENT OUTCOMES FOR NATIONAL BOARD CERTIFIED TEACHERS AND NON-NATIONAL BOARD CERTIFIED TEACHERS

A Dissertation Presented

to

The Faculty of the School of Education

Learning and Instruction Department

In Partial Fulfillment

of the Requirements of the Degree of

Doctor of Education

by

Wendy Hacke, NBCT

San Francisco

May 2010

ABSTRACT

Current quantitative research on the effectiveness of the National Board certification has resulted in contradictory findings. Consequently, this meta-analysis synthesized the results of 12 studies on National Board Certification and student achievement. Of those, 9 studies were used to calculate effect sizes for reading, and all 12 were used to calculate the effect sizes for mathematics. On average, students of National Board Certified teachers had higher achievement scores; however, the effect sizes were small. Similarly, there was no difference in the student outcomes for NBCTs in either mathematics or reading. Due to the limitations of the meta-analysis, analyzing study characteristics as possible moderator variables using tests of homogeneity and analog to the analysis of variance did not lead to finding any variables that accounted for difference in study results. The results, however, did provide direction for future research in the area of National Board Certification.

This dissertation, written under the direction of the candidate’s dissertation committee and approved by the members of the committee, has been presented to and accepted by the Faculty of the School of Education in partial fulfillment of the requirements for the degree of Doctor of Education. The content and research methodologies presented in this work represent the work of the candidate alone.

Wendy Hacke, NBCT______________ 7/21/2010
Candidate Date

Dissertation Committee

Dr. Lanna Andrews 7/21/2010
Chairperson

Dr. Patricia Busk 7/21/2010

Dr. Caryl Hodges 7/21/2010

DEDICATION

I would like to dedicate this research to my family, who has supported all of my educational endeavors. To my husband, the love that God sent me to treasure, who has always encouraged me to fulfill my dreams and achieve my aspirations. To my son, my shining star, who proofread papers late into the night even though he was working full time and trying to earn his own Jurist Doctorate. To my daughter, my most prized possession, who used her burgeoning library skills to help me research papers as well as my dissertation while she worked on earning a Masters Degree in Library Science. To my daughter-in-law, the latest addition to our family, who has encouraged and prayed for me as I went through my dissertation journey. Finally, to my church family, who prayed for strength and wisdom for me as I traversed the bumpy, twisted road of completing my dissertation. I have been truly blessed to have loving and supportive family and close friends, like Sharon, who sacrificed time in order for this dissertation to be completed.

To Jesus Christ, my savior… May He bless us all!

ACKNOWLEDGEMENTS

My dissertation “journey” has been an extremely challenging and invigorating experience. The journey could not have been accomplished without guidance, support, and encouragement from many special individuals. Those special people, that God blessed my life with, through out the years of working toward my doctorate, aided in making my dream a reality. All the encouragement and effort from each individual has been a blessing and I want to say thank you for helping me through the process.

To offer a mere thank you to acknowledge my appreciation for the consistent help and encouragement afford me by my dissertation chair, Dr. Lanna Andrews insufficient. Dr. Andrews who willingly agreed to chair my dissertation committee and spent a good deal of time getting me through the process of completing a meta-analysis. From our first encounter when I was a doctoral fellow to our last conversation about my dissertation you have been a source of constant guidance and motivation-you are the best! Dr. Patricia Busk, through all six statistics classes and my dissertation you have always pushed my thinking and required me to write clearly, so that the reader can understand my thinking. Appreciation is also extended to Caryl Hodges. Thank you for providing welcomed feedback and for graciously agreeing to serve on my committee.

Appreciation also goes to special individuals who have supported me professionally and personally over the course of my career as an educator. Neva Hofemann, thank you for all you have done to encourage my growth as a professional. From the first class in *Good Teaching Practices,* to working for you at the University I have appreciated your mentoring and your guidance in my professional growth. I will always admire your vision, courage, and strength, and I truly appreciate the opportunities you have given me to grow. Mary Howland you were truly a God send. I enjoyed our carpool conversations, your help with statistics and my dissertation, and co-teaching in the Tier I program. Reylene Potter you were truly helpful with coding studies for the meta-analysis. Finally, I wanted to thank Pamela Root for reading a chapter and checking for citations that needed to be included in the reference section.

Ultimately, for placing these treasured people in my life, I thank God who has strengthened me, blessed me, and provided the path that lead to reaching my dissertation destination (Mark 10:27)

TABLE OF CONTENTS

TABLE OF CONTENTS Continued

CHAPTER Page

TABLE OF CONTENTS Continued

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Statement of the Problem

From the time *A Nation at Risk: The Imperative of Education Reform* was published in 1983 through the reauthorization of the federal Elementary and Secondary Education Act, known as No Child Left Behind (NCLB), in 2001 (20 U.S.C. § 6301 *et seq.*) to the present, effective teaching has remained squarely in the middle of state and national agendas as the key to successful student learning. The NCLB legislation defines teachers as highly qualified if they hold a bachelor's degree from a 4-year institution, hold a full state teaching credential, and demonstrate competence in their subject area(s). This definition reflects the view that state teacher credentialing alone does not ensure teacher quality. Research also has explored this issue and has shown that only very knowledgeable and skillful teachers who are able to respond appropriately to students' needs have the ability to challenge and support students' academic growth (Kennedy, 2006). In 1987, the National Board for Professional Teaching Standards (NBPTS) grew out of the research on effective teachers to become part of the infrastructure that provides assessment of the teaching methods, practice, and the processes for building knowledge from practice. The task of the Board was to develop a national, voluntary certification process that measures teacher quality and identifies the most effective teachers (Hakel, Koenig, & Elliott, 2008).

Because teacher effectiveness is intended to result in student achievement, National Board Certified Teachers (NBCTs) were surveyed in the Fall of 2001 (NBPTS, 2001) regarding student achievement outcomes. Sixty-nine percent reported that their students’ engagement, motivation, and achievement increased when the teachers achieved certification. The problem is that, although there is an ever-growing body of research to analyze NBCTs’ professional development experiences and their relationship to student achievement (Vandevoort, 2004), comparisons of studies of student outcomes have produced a copious assortment of positive, null, and negative results (Amrein-Beardsley & Berliner, 2004; Clotfelter, Ladd, & Vigdor, 2006; Harris & Sass, 2007; Sanders, Ashton, & Wright, 2005; Vandevoort, 2004). Reviews of these studies have failed to answer the question of whether National Board Certification identifies effective teachers who increase student learning outcomes (Hakel, Koenig, & Elliott, 2008; Holland, 2006; Leef, 2003; Predrosky, 2001; Stone, 2002).

The need for further systematic work to investigate the factors associated with the variability of empirical studies of the relationship between National Board Certification (NBC) and student achievement lends itself to the meta-analytic method of research (Cooper & Hedges, 1993). The comprehensive integration of findings on the topic of NBC would permit the examination of meaningful effects and relationships in order to bring coherence and perspective to the problem.

Meta-analysis takes the results of numerous studies of the same research question and combines them into a single analysis. The purpose of this study was to conduct a meta-analysis in order to aggregate the research findings of empirical studies that investigated the relationship between National Board Certification (NBC) status and student achievement, and by analyzing moderator variables. Although a quantitative meta-analysis has the potential to address the limitations of past research on the subject of National Board Certified Teachers’ (NBCTs) effectiveness in increasing student learning outcomes, it has not been used as a method to explore the topic. In order to resolve the issue of addressing the limitations of past research, the current study used the meta-analytic model to aggregate the findings of 12 studies that investigated the academic gains of students who were taught by two groups of teachers: (a) board certified teachers and (b) teachers who are not certified.

Purpose of the Study

The purpose of this study was to conduct a meta-analysis in order to generate new evidence by analyzing moderator variables and examining the aggregated research findings of studies that explored the relationship between certification status and student achievement. The descriptive data of 12 studies on NBC and student achievement were examined to assess the comparative teaching outcomes of NBCTs and nonNBCTs for the purpose of creating generalizations. In addition, outcomes for these two categories of teachers were assessed across school levels and subject matter. Finally, study characteristics associated with differences in effect sizes were identified in order to search for influences on previous findings in order to resolve conflicts in the literature.

Conceptual Framework

Several aspects of research used to examine the effects of teachers and teaching on student achievement provide the conceptual basis for this study. Most studies of NBC have used a conceptual model developed by Bond, Smith, Baker, and Hattie (2000). In developing a construct of effective teaching, Bond and his fellow researchers conducted a comprehensive review of the literature and identified 15 key dimensions of teaching, which fell into three broad areas that were the focus of their study: Comparative Teaching Practices, Comparative Teaching Outcomes, and Comparative Professional Activities.

Bond and his colleagues then developed protocols to evaluate board certified and nonboard certified teachers on these dimensions. Using (a) reviews of teacher assignments and student work, (b) student interviews and questionnaires, and (c) classroom observations that evaluated student self-efficacy and motivation as well as classroom climate and environment, the researchers found that board certified teachers performed higher on all 15 dimensions than candidates who were unsuccessful in their attempt at certification and were considerably higher on 13 of these dimensions of teacher expertise. The other two dimensions were still higher for NBCTs, but the difference was not statistically significant.

Although Bond’s conceptual framework lends itself to the qualitative studies of NBC, where observations, student samples, and surveys are used to investigate teacher effectiveness, a connection between effective teaching and teacher effects is not provided. Ding and Sherman (2006) argued that teaching effectiveness is conceptually different from teacher effect and that the misuse of these two concepts influences efforts to improve student achievement through effective teaching. Ding and Sherman defined teacher effects as salary, gender, and years of teaching, which is different from teacher effectiveness, generally characterized by instructional factors that influence the amount of gain students show on standard achievement tests. These characterizations can be operationalized in different ways, but without a clear definition of the concept of teacher effectiveness its relationship with other factors cannot be verified empirically (Ding & Sherman, 2006; Goe, Bell, & Little 2008; Odden, Borman, &. Fermanich, 2004; Veldman & Brophy, 1974).

In their research of the literature, Ding and Sherman (2006) found that the missing element in studies of teacher effect on student achievement gains was an in-depth definition of teacher effectiveness. In light of the need for a broader and more comprehensive definition of teacher effectiveness, they developed a multilevel educational model based on the previous work of Odden et al., 2004. Both sets of researchers argued that the use of test score data to estimate teacher effectiveness requires the acknowledgement of the nested nature of school, teacher, and student factors. The difference is that Odden et al. did not include a variable on teacher effectiveness. The model depicted in Figure 1 (see page 6), is based on Ding and Sherman’s model that differentiates teacher factors from teacher effectiveness and is used in this study as a framework to examine the relationship between NBC and student achievement gains. Previous studies of NBC as an indicator of effective teaching based on student achievement were void of a definition; therefore, defining teacher effectiveness as more than teacher effects provided the guiding framework for this meta-analysis. Specifically, the definition was used both in the selection of literature to discuss in chapter II and for determining moderator variables to be coded for analysis.

If board certified teachers do have a greater influence on student learning and growth than nonboard certified teachers after controlling for individual, school, and teacher factors, it would be suggestive of a verifiable difference in instructional effectiveness. This premise is why the current study used meta-analytic procedures to explore the relationship between NBC and student achievement. If board certification is an effective signal of teaching ability, then nontrivial effects on students’ achievement growth in an academic year should be evident. Several studies (Hill, 2005; Porter, 2009; Sartawi, 2009) found that the use of Ding and Sherman’s (2006) model as a construct was

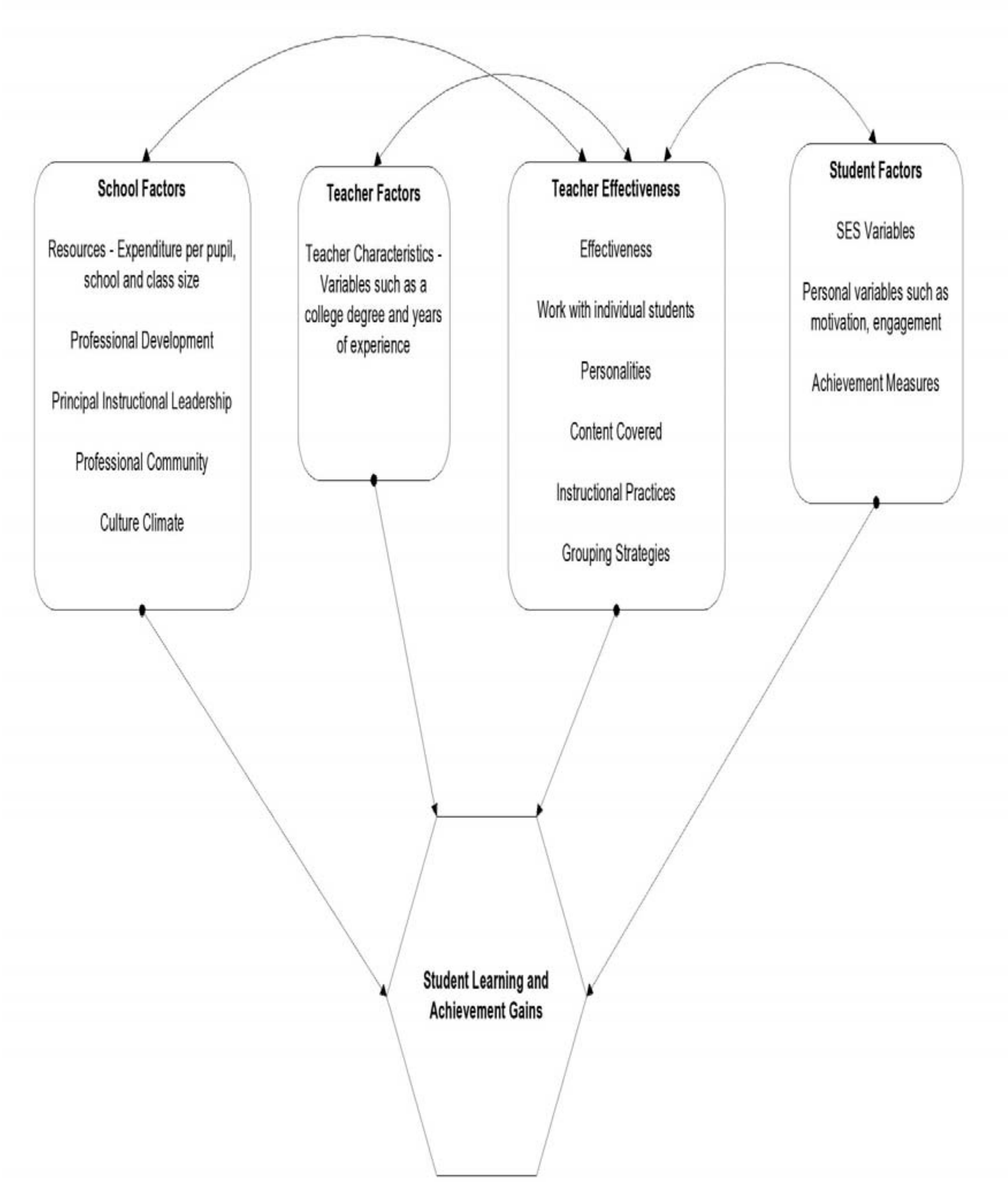

School Factors
Resources - Expenditure per pupil, school and class size
Professional Development
Principal Instructional Leadership
Professional Community
Culture Climate
Teacher Factors
Teacher Characteristics - Variables such as a college degree and years of experience
Teacher Effectiveness
Effectiveness
Work with individual students
Personalities
Content Covered
Instructional Practices
Grouping Strategies
Student Factors
SES Variables
Personal variables such as motivation, engagement
Achievement Measures
Student Learning and Achievement Gains

Figure 1. Multilevel Education Model of Factors Influencing Student Achievement. Based on Ding and Sherman's (2006) multilevel dynamic education model of school factors, teacher factors, student factors, and teacher effectiveness on student learning.

helpful in identifying teachers who achieved a level of competence in the domains of knowledge, skills, and judgments. The idea was that if NBCTs do help students achieve higher academic achievement, then the effect sizes found in the current study would reflect the relative magnitude of classroom-to-classroom differences, as defined in Ding and Sherman’s model, between board certified and nonboard certified teachers. The findings of these studies support the use of the multilevel dynamic education model as a construct and, therefore, provide the rationale for the choice of teacher effectiveness as the criterion variable of the current study.

Since 1990, research has demonstrated that teacher effectiveness is the most dominant factor effecting student academic achievement (Rowan, Chiang, & Miller, 1997; Sanders & Rivers, 1996; Wright, Horn, & Sanders, 1997). It is imperative that teachers improve their professional skills and increase their effectiveness. To accomplish the goal of increasing skills and effectiveness, the nation must ensure that all teachers have access to high quality professional development (National Commission on Teaching and America’s Future, 1996). Existing research, however, does not provide documentation that the NBC process, as a method of professional development, enhances teachers’ skills at improving students’ annual achievement gains (Hakel et al., 2008). To address this research-to-practice gap, the current meta-analysis used the multilevel dynamic education model for comparing student outcomes for students of NBCTs and nonNBCTs. Given this model, if there are classroom-to-classroom differences between board certified teachers and nonboard certified teachers, there should be differences in student achievement outcomes.

For this reason, the current meta-analysis explored differences in student achievement growth in mathematics and reading and how the factors included in the multilevel dynamic education model moderates the differences (Ding & Sherman, 2006). The study also investigated the possibility that there would be less variance in academic achievement gains when NBCTs teach different academic subjects, as a result of effective teaching (Harris & Sass, 2007). Finally, the study examined the likelihood that academic gains would be similar across 3rd through 12th grade for students in both reading and mathematics resulting from the instructional practices of NBCTs (Smith, Gordon, Colby, & Wang, 2005).

If NBCTs possess the necessary knowledge, skills, classroom practices, dispositions, and judgments that the National Board for Professional Teaching Standards (NBPTS) claims, then variability of student achievement outcomes across empirical studies would be associated with the relationship between the two variables. More specifically, the variety of factors that influence student achievement would be the variables that account for the variation in the magnitude of the difference between certification status and student achievement gains regardless of grade level or subjects taught.

Research Questions

The majority of quantitative studies that estimate the effects of NBC focus on student achievement, as measured by annual standardized tests. This quantitative synthesis, therefore, investigated these results by addressing the following questions:

1. What is the effect on student achievement in mathematics and reading for students taught by teachers with National Board Certification (NBC) when compared with students taught by teachers without NBC?

2. What is the difference in the effect size of reading and mathematics assessments for students taught by NBCTs when compared with students taught by nonNBCTs?
3. To what extent do study features moderate the relationship between certification status and student academic achievement?

Background and Need

David C. Berliner, in his article *In Pursuit of the Expert Pedagogue* (1986), suggested that finding an effective teacher first requires distinguishing criteria for identifying expertise in pedagogy. According to the National Board for Professional Teaching Standards (NBPTS), these criteria have been met by the teachers certified through their process. At its inception in 1987, the Board represented the cutting edge of developing professional expertise in the field of teaching. Created as an outgrowth of recommendations of the Task Force on Teaching, which was established to address concerns raised by a federal report titled *A Nation at Risk: The Imperative of Education Reform* (1983), the Board's mission was to develop rigorous standards for accomplished teaching. In view of the fact that the impetus for the development of these standards came from the idea that effective teachers are necessary for student learning (Shulman, 2000), the National Board believed it was necessary to codify the knowledge of specific expertise required by experienced teachers. This codification then was used to create assessment criteria designed to appraise multiple dimensions of effective teaching, including knowledge of discipline and of how to diagnose and treat students with various learning needs (Glazer & Hannafin, 2006). Therefore, the NBPTS has become part of the infrastructure that provides methods, practice, and processes for building knowledge from practice. Nearly 3% of the 3.7 million teachers currently eligible for certification have achieved National Board Certification (NBC) since the voluntary process began in 1994 (NBPTS, 2008).

National Board Certification is an advanced teaching credential with certification signifying that the holder has met high and rigorous standards for what accomplished teachers should know and be able to do (Hakel et al., 2008; Ingvarson & Hattie, 2008; Silver, Mesa, Morris, Star, & Benken, 2009). From the time it was established, the Board's mission was to develop rigorous standards for accomplished teaching based on theory and research centered on specific goals for student learning. By codifying teacher knowledge, skills, and dispositions that account for accomplished practice, the NBPTS created assessment criteria that could be used to judge teacher practice that are content specific and emerged from consensus among practitioners rather than solely from empirical research (Hakel et al., 2008; Ingvarson & Hattie, 2009; Smith, 2004). The codification process involves a committee, the majority of whom are teachers in the area of certification, researchers, higher education representatives, and experts in the field to be assessed (NBPTS, 2002). The members are chosen to reflect gender, ethnicity, geographical, and teaching context diversity. The task of the committee is to look at the standards of professional organizations as well as theory and research in the field for which they are codifying the knowledge, skills, and dispositions necessary to demonstrate accomplished teaching (Camp, 2007).

The committee bases its decisions on the original codification of five core propositions articulated in the NBPTS (1989) publication *What Teachers Should Know and Be Able To Do*. These five principles that reflect the original committee's vision of accomplished teaching are

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

Situated in actual practice, the experiential certification process has two main components. First, there is the portfolio, which is a time-consuming, rigorous, and at times tedious process embedded in the day-to-day practice of teachers (Lustick & Sykes, 2006). The portfolio requires participants to think systematically about their practice and learn from experience, using student work samples and videos of actual teaching sessions. It includes three classroom-based entries, two that require submission of videotaped classroom instruction and one that exhibits work in the community, with families, and colleagues and the larger profession (Goe et al., 2008). Second, there is the 4-hour Assessment Center exam that is administered in six parts to assess content knowledge and the use of effective student assessment. Both parts of the NBC process encourage collegiality through collaboration and discourse and lead to transformative learning as teaching beliefs and practices change (Cohen & Rice, 2005).

Influenced by the Institute for Research on Teaching (IRT) and later the Teacher Assessment Development Project (TAP), which were both led by Lee Shulman, the chief architect on the NBPTS, the rigors of the NBC process are grounded in the construct of knowledge being acquired from practice (Hakel et al., 2008). Based on observations of assessments used in other professions, Shulman and his colleagues created a process that requires candidates to spend more than 400 hours assembling direct evidence of their expertise in content knowledge, meeting individual student needs, and using assessment to plan instruction (Harris & Sass, 2007). The most important part of this collection process is the fact that teachers are required to write detailed analyses of their teaching as well as detailed reflective analyses of their teaching practices. This self-reflection on teaching practices starts by teachers critically thinking about questioning, analyzing, and re-analyzing their teaching behaviors and choices and ends with them considering alternative actions and reactions and anticipating consequences (Weglinsky, 2004). The self-reflection required by the NBC process also provides teachers with the opportunity to make sense of and integrate their experiences into the process.

The online portion of the National Board Certification process is 40% of the total score and is administered at over 300 computer-based testing centers across the United States. Candidates have up to 30 minutes to respond to each of six exercises that demonstrate mainly their content knowledge, with a limited number of questions also covering pedagogical strategies. These assessment exercises were developed and designed by practicing professionals in the certificate area in order to gage the candidate's fundamental content knowledge that supports effective instruction on a daily basis (Hakel et al., 2008; Silver et al., 2009). Through their responses, candidates are expected to demonstrate knowledge of appropriate content across the full developmental age range included in their certificate area.

Once the year-long process culminates in the portfolio and submission of test responses, both the portfolio and assessment-center responses are scored by teachers in the field covered by the certificate area being assessed. These volunteers may or may not be board certified, but they all are provided extensive training on how to score entries. Scores reflect the candidate’s depth of understanding of subject matter concepts and processes, as well as the accurate identification of a student’s misconception or difficulty with the content of instruction. Scorers also look for well-developed instructional strategies or learning experiences that clearly are linked to real-world applications and appropriately address a child’s needs. The developmental appropriateness of the choice of materials to teach a concept and a rationale for the choice of these materials also are scored (Ingvarson & Hattie, 2008).

Each entry is weighted with the highest weight placed on the classroom-based entries, with each valued at 16% of the total score. Twelve percent is the weight given the documented accomplishment entry. The six assessment center exercises are each weighted 6.67%. The combined weighted scores produce a scaled score of 1 to 400 with 275 being the set score for certification. Because there is a 50% certification rate for first time applicant, if a candidate does not reach the set score of 275 he or she may bank any score that is over 2.75 and reattempt any entry or exercise that is less than 2.75. Candidates who choose to retake a portion of the assessment pay an additional fee, and they have 2 years to raise their scaled score to at least 275. Once a candidate certifies, their certification is good for 10 years and can be renewed by completing a shorter version of the portfolio process (Hakel et al., 2008).

Current quantitative research findings on the NBC’s effectiveness show conflicting results. The sharpest differences involve the question of whether teachers certified by the NBPTS are more effective than teachers without that status (Clotfelter, Ladd, & Vigdor, 2007). Two examples of the disparity between the results are studies completed by Cavalluzza (2004) and Stone (2002). On the one hand, Cavalluzza (2004) found that having a National Board Certified Teacher (NBCT) in the classroom had the effect of adding over a month's worth of education to the school year. Stone's (2002) research, on the other hand, revealed no difference between the achievement outcomes of students with and without NBCTs. To date, only one effort has been made to synthesize the literature on NBC and look for empirical evidence to inform the debate.

The U.S. Congress commissioned a study that was conducted by the National Research Council (NRC; Hakel et al., 2008) to evaluate the effect of NBC. The NRC used only seven quantitative studies to yield findings regarding the relationship between board certification and student achievement. The small numbers of studies included in their review were used to draw conclusions on the effectiveness of NBCT in terms of student learning outcomes but no statistical analysis was conducted to confirm their findings. Furthermore, in reviewing studies that examined the relationship between board certification and student achievement, NRC (Hakel et al., 2008) found that only one study (Cantrell, Fullerton, Kane, & Staiger, 2007) of the seven they reviewed did not have interpretation issues. Four of the studies, Cavaluzzo (2004), Clotfelter et al. (2006), Goldhaber and Anthony (2007), and Sass (2007) had standard errors that did not account for nesting. This aspect of the study features could explain why the studies yielded different results. In another of the studies, Sanders, Ashton, and Wright (2005) analyzed grade levels separately, which may have reduced power because the sample size was too small. The seventh study was a second study by Clotfelter et al. (2006) and it did not focus primarily on the evaluation of the effectiveness of NBCTs.

McCaffey and Rivkin (2007), who conducted another report for the NRC, further found that some fixed effects estimates might be biased because they may fail to account for confounding factors. McCaffrey and Rivkin also found that although a fixed effect model can eliminate unobservable cross-sectional individual differences that affect achievement many researchers did not address them in their statistical model. They also found that those researchers who did may have had their estimates attenuated by inability of the model to address problems caused by the purposeful sorting of students and teachers within schools. Finally, McCaffrey and Rivkin suggested that the difficulty of separating the causal effects of NBC from other cofounding factors might bias the NBC effects and cause the disparity in the results of various studies.

To solve the above problems, some studies combined school and student-fixed effects but computational concerns arose because the numbers of students and teachers that can be used in the study are limited by the fact that only students who have been in a school for 2 or more years can contribute to the estimates (Burks & Sass, 2008). Additionally, school-fixed effects estimates may be biased by time, varying aspects of the school, or student quality (Jargowsky & El Komi, 2009).

These issues lent themselves to the need for a meta-analysis, which combined results for greater statistical power and aggregated data by school level (elementary school, middle school, and high school), to compare the student achievement of NBCTs and nonNBCTs. Additionally, there was a need to use the educational model developed by Ding and Sherman (2006) as a framework for examining the influence of school, student, and teacher effects on student academic gains in order to explain the relationship among concepts. Finally, there was a need to explore the influence of methodology and methodological quality on study outcomes in comparing the conflicting results of studies included in this meta-analysis.

Significance of the Study

This meta-analysis has implications for school systems seeking to improve teacher effectiveness and student learning outcomes by addressing the dilemma of inconsistent findings of past research. The current study is important for three reasons. First, it should provide more consistent and valid answers to the question of whether National Board Certification identifies effective teachers who increase student learning outcomes by considering the influence of moderator variables on final outcomes. The disparity in the results of previous research that examined the relationship between NBC and student achievement using test scores lends itself to the need for studying subject and student variability, as well as school and classroom effects as important contributors to the outcome of an investigation (Ding & Sherman, 2006; Olejnik, 1988).

Second, accruing information from a number of primary studies aids both in accumulating evidence and generating new evidence that can inform the debate regarding NBC as a predictor of increased student learning outcomes and identify central issues, ideas, and theories for future research. By exploring the efficacy of the NBC system, this meta-analysis addressed the disparity in results of previous studies regarding the effect of certification on student achievement. State standards, accountability systems, and the federal No Child Left Behind Act have placed demands on schools to improve teacher effectiveness and student achievement (Heneman, Milanowski, Kimball, & Odden, 2006). As part of that accountability system, virtually all states have constructed subject-matter content standards and methods for assessing the mastery of those standards. Their development has underscored the importance of considering specific aspects of teacher knowledge and the application of that knowledge, which has important ramifications for policymakers, educators, students, and society. A system that can distinguish those teachers who can facilitate greater levels of student achievement would ensure the creation and maintenance of a high quality teaching force with the competencies to help children learn.

Finally, this study contributes to the national conversation on teacher effectiveness. Since the 1970s, teacher effectiveness has been defined as the amount of gain students show on standard achievement tests (Ding & Sherman, 2006; Goe et al., 2008; Odden et al., 2004; Veldman & Brophy, 1974). Although few would argue that the gains students show on standardized achievement tests is the best method or the only method of measuring teacher effectiveness because it is increasingly used for this purpose, the studies analyzed in this meta-analysis focused explicitly on examining the differential effect of teachers on student achievement scores. This narrow definition ignores the evidence that the effects given by a variable in a particular study depend on whether other variables also possibly measure aspects of effectiveness (Darling-Hammond & Youngs, 2002; Ding & Sherman, 2006). For this reason, the present study focused analysis efforts on multiple factors (moderator variables) that influence study outcomes with respect to the effect of NBC. It also was important to focus on these multiple factors because they are prevalent in the current education policy landscape and are areas in which stakeholders and critics alike indicate a need for more empirical evidence.

Definition of Terms

Given the various uses and interpretations of educational terminology, the following section uses definitions that are most likely to be encountered in the literature to delineate terms that apply to this meta-analysis.

Analog to ANOVA

In the analog to analysis of variance (ANOVA), the Q statistic from the analysis of variance is subdivided into $Q_{between}$, representing the variance in effect sizes accounted for by moderator variable, and a Q_{within}, representing within group error. When the $Q_{between}$ is statistically significant and the Q_{within} is not statistically significant, the moderator variable successfully accounts for the variability in effect sizes (Lipsey & Wilson, 2001).

Certification

According to the National Research Council of the National Academies (Hakel et al., 2008) certification is a voluntary means of establishing that an individual has mastered specific sets of advanced skills that that come with expertise over time.

Contextual Factors

Odden et al. (2004) define contextual factors as socioeconomic characteristics of the classroom including poverty and race variables as well as student grouping practices. This definition is further expanded by Jargowsky and El Komi (2009) who included neighborhood and peers.

Effect Size

According to Lipsey and Wilson (2001), critical quantitative information from relevant study findings are encoded on a statistic called an effect size. Specifically, an effect size is a statistical concept that measures the strength of the relationship between two variables. Different effect-size statistics are required for different study findings in order to produce statistical standardization (Lipsey & Wilson, 2001). Because two groups are being contrasted in this meta-analysis, the standardized mean difference effect size statistic was calculated; however, this statistic has been shown to be upwardly biased, so Hedge's unbiased effect size (Hedge's *g*) was calculated. For this study, the effect-size measurements are used to compare the magnitude of differences between student achievement for students of nonNBCTs and NBCTs.

Licensure or Credentialing

According to the Certification Board for music therapists licensure that is synonymous with credentialing, refers to the laws that regulate a given occupation. Its purpose is essentially twofold: (a) title protection, that is, the prevention of unqualified individuals utilizing the given title, and (b) scope of practice that is defining the specific tasks that constitute the practice of the given occupation. Certification, alternatively, is a nonstatutory process whereby an accrediting body grants recognition to an individual for having met preinvestigated professional qualifications (Oliver, 2010).

Moderator Variable

Durlak (1995) defines a moderator variable as a study characteristic that accounts for significant variability of effect sizes among reviewed studies.

National Board Certification

National Board Certification specifically refers the to advanced skills that have been codified by the National Board for Professional Teaching Standards as part of their voluntary process of identifying effective teachers (NBPTS, 2005).

National Board Certified Teacher

Teachers who achieve the National Board standards are referred to as National

Board Certified Teachers (NBCTs) (Bundy, 2006).

Non-National Board Certified Teacher

A nonboard certified teacher (nonNBCT) is a teacher licensed by the state of employment to teach in that state who has not been certified by the National Board for Professional Teaching Standards (Benigno, 2005).

School Factors

Odden et al. (2004) define school factors as a school's capacity to support teachers in providing effective instruction. Ding and Sherman (2006) expand this definition by identifying resources, professional development, instructional leadership, professional community, and cultural climate as factors that support effective instruction.

Student Achievement

Student achievement is defined as end-of-year or end-of-instruction test score gains on standardized tests in reading and mathematics (Cabezas, 2006).

Student Factors

Student factors are defined by Odden et al. (2004) as actions or dispositions of individual students that impact their own learning. Ding and Sherman (2006) include socioeconomic variables, motivation, engagement, and achievement measures as factors for this category in their multilevel dynamic education model.

Teacher Factors

Ding and Sherman (2006) defined teacher factors in their multilevel dynamic education model as teacher characteristics such as college degrees and years of experience. Other researchers, such as Odden et al. (2004) include teacher licensure, college major, verbal ability, and coursework.

Teacher Effectiveness

Ding and Sherman (2006) and other researchers (Goe et al. 2008; Odden & Borman, 2004; Veldman & Brophy, 1974) defined teacher effectiveness as the effect a teacher has on the amount of gain student's show on standard achievement tests.

Ding and Sherman, (2006) and other researchers (Goe et al., 2008; Odden et al., 2004) defined teacher factors as teacher licensure, years of teaching, major of undergraduate study, American College Testing (ACT) and Scholastic Aptitude Test scores (SAT), degrees obtained, and verbal ability.

Test of Homogeneity

A *Q* test is used to evaluate the computed effect sizes for homogeneity (de Liz & Strauss, 2005). Testing for homogeneity before estimating a mean effect size is conducted to learn whether the effect sizes share a common population effect size. (Kim, 2000, de Liz & Strauss, 2005). When each effect size does not estimate a common population mean, the difference may be associated with different study characteristics (Lipsey & Wilson, 2001).

Summary

Considered the key to effective learning, effective teaching has been the focus of both legislation and research. Research that has shown that only the most knowledgeable and skillful teachers have the ability to meet the needs of students and support their academic growth has also lead to studies on the effectiveness of National Board Certification. Current quantitative research on the effectiveness of the National Board Certification has resulted in contradictory findings. Consequently, this meta-analysis synthesized the results of 12 studies on National Board Certification and student achievement. The results of this study add to the literature on the question of whether National Board Certification identifies effective teachers who increase student learning outcomes. It also contributes to the national conversation on teacher effectiveness. Finally, the results also may be used as indicators for future research.

A review of the literature follows in chapter II, which builds on the background and need for the study and examines the research on variable that may moderate the findings. These include teacher, student, and contextual factors that research has shown to contribute to variation in study results. Other factors discussed because of their possible contribution to differences in research outcomes are methodological quality and the use of different research methods in studies examining the effectiveness of National Board Certified Teachers. Methodology, findings, and a discussion of the findings follow in chapters 3, 4, and 5, respectively.

CHAPTER II

REVIEW OF THE LITERATURE

In the introductory chapter, the argument was presented that there is conflicting evidence regarding the effectiveness of National Board Certified Teachers (NBCT) when compared with nonboard certified teachers, especially as measured by student achievement. This discord has made it difficult to acknowledge the effect of the certification process on teacher practices and ultimately on student learning. Therefore, it is essential to investigate the factors associated with the variability of empirical studies of the relationship between National Board Certification (NBC) and student achievement, which is the goal of the present meta-analysis. The following review of literature provides a framework for this meta-analysis.

The first two sections of this chapter build on the argument in the introductory chapter of this dissertation. The first section, *Research Defining Teacher Effectiveness*, provides the background and need for the conceptual framework of the study. The following section, *Models and Measures Used to Study Teacher Effectiveness*, presents a critical look at research models that have been used to assess effective teaching and the current use of student gain scores to measure teacher effectiveness. Next, given that the purpose of this meta-analysis is to generate new evidence by analyzing moderator variables that might explain the variability in findings regarding the relationship between certification status and student achievement, the remaining section, *Various Influences on Study Outcomes,* contains the research on potential moderator variables used in this meta-analysis.

Research Defining Teacher Effectiveness

This section focuses on the history and definition of teacher effectiveness as well as links to NBC research. A review of the literature on teacher effectiveness provides a context for current attempts to correlate certification status with student achievement and addresses the ambiguity in terms that provides the rationale for the conceptual framework. This overview of the historical background of research on effective teaching and the definitions that have influenced the research is vital for understanding the variance in the research findings that have persisted since the first investigations on the topic. It is also important for discerning the relevant issues in evaluating the value of NBC as a signal of effective teaching.

Historical Perspective

Each era of study has contributed new understandings of teacher factors and the effect they have on student achievement. Since the early 1920s, teacher effectiveness has taken a prominent place amid the quantitative research in the field of education (Doyle, 2010; Hill, 1921). By midtwentieth century most of the research conducted on teacher effectiveness had focused on teacher characteristics and behaviors (Medley, 1977b). Much of the research that was done to evaluate teacher effectiveness during this period centered on surveys of students who provided lists of characteristics and behaviors of teachers they believed were most effective. Although the lists were extensive, they were neither empirically supported nor linked to student outcomes.

The focus shifted in the 1950s with the formation of the American Educational Research Association (AERA) committee on *Criteria of Teacher Effectiveness* (Doyle, 2010). Their publication of the *Handbook of Research on Teaching* (Gage, 1972; Good,

1979) came at a time when attention was focused on educational outcomes and the act of teaching. Acts of teaching were less related to teacher characteristics and more on the process of instruction. At this time, research focused on polar opposites like formal and informal assessment and progressive and traditional teaching methods (Campbell, Kyriakides, Muijs, & Robinson, 2003).

The boon years of the 1960s brought renewed interest in teacher behaviors and ushered in two decades of process-product studies in which specific teacher actions were connected directly to student outcomes (Galton, 1987; Palmer, Stough, Burdenski, & Gonzales, 2001; Shulman 1986a; Stephens, 2003). It was during this period that the Coleman Report (Coleman, 1966) was published, which was designed by the Office of Education to investigate what factors played a role in student achievement. The findings of the study were reported to Congress as part of the Civil Rights Act of 1964. After studying over 570,000 students and 60,000 teachers, the Coleman Report minimized the importance of the teacher, asserting that moderators like family background and socioeconomic status were the major causal variables affecting difference in achievement (Benigno, 2005). Notwithstanding this study’s pessimistic view of the value of teachers, researchers continued to focus on teacher effectiveness as a major factor in student achievement (Benigno, 2005; Blanton et al., 2003; Hanushek, 2004; Nye, Konstantopoulos, & Hedges, 2004).

From the effective schools movement of the late 1980s to the present value-added studies, the focus has remained on student outcomes as the product. Research, however, swung back to teacher characteristics with Shulman’s (1987) publication of *Knowledge and Teaching: Foundations of the New Reform* and the release of the findings in a *Nation* *at Risk: The Imperative for Educational Reform* (National Commission on Excellence in Education, 1983). These publications ushered in the effective schools movement, the standards movement, and a research focus on teachers' subject matter and pedagogical knowledge and beliefs (Blanton et al., 2003; Campbell et al., 2003; Goe, 2007). They also laid the groundwork for establishing the National Board for Professional Teaching (NBPTS), which developed teaching standards based on research into the effectiveness of teacher knowledge and beliefs on student achievement. The impetus for the development of these standards came from the research findings of these studies, which demonstrated that quality teachers are necessary for student learning (Benigno, 2005; Campbell et al., 2003),

Modern research on teacher effectiveness has evolved since the Coleman Report (Coleman, 1966). The final shift in teacher effectiveness research came with the American Research Council (1999) finding that the most influential element for students' successfully learning is the quality of the teacher. In part the shift resulted from earlier researchers not having the methodology for accurately measuring the variables (Harris & Sass, 2009; Haycock, 1998; Medley, 1977a). Since the mid 1980s, administrative databases that track individual student achievement over time have altered radically how educational research is conducted. At the time of this meta-analysis, due to this improved technology and research methodology, there is an extensive body of research that has demonstrated that effective teaching is the most dominant factor in affecting student academic achievement (Darling-Hammond & Youngs, 2002; Hanuskek, 2005;Rowan, Chiang, & Miller, 1997; Sanders & Rivers, 1996; Wright, 1997).

One such methodological improvement developed by William Sanders is the value-added approach that uses prior students' performance to annually classify teachers as effective or ineffective (Cunningham & Stone, 2005). Made possible by technological advances in the maintenance of data, value-added research investigates the cumulative effects of teachers on student achievement over grade levels (Sanders & Rivers, 1996). These studies have overwhelmingly demonstrated that teacher effectiveness is the most dominate factor in student achievement (Ballou, Sanders, & Wright 2004; Kane & Staiger, 2008b; McGraffrey et al., 2004; Rothstein, 2009; Sanders & Rivers, 1996; Stone, 2002).

Varied Definitions of Teacher Effectiveness

This subsection presents the complexities of defining and describing teacher effectiveness in order to measure it. The need for an explicit meaningful definition underlies the conceptual framework for using standardized achievement scores to assess teacher effectiveness by evaluating their students' outcomes.

With research confirming a direct link between teacher competency and student outcomes, it is clear that the identification of effective teachers is crucial for the achievement of all students. It has become so clear that current Federal Legislation holds states accountable for ensuring that students have "highly qualified teachers." Under the guidelines of No Child Left Behind (NCLB), a teacher is highly qualified if he or she holds a bachelors degree, certification or licensure to teach in their state of employment, and has proven competence in the subjects that she or he teaches (U.S. State Department of Education, 2004). This definition is interpretable by states, which differ greatly in their specifications for meeting the federal requirements. The wide variation in the application of the NCLB requirements is evident in research, which has demonstrated that the designation of “highly qualified teachers” does not translate into the same level of academic achievement for all students (Angle, 2006; Goe, 2007; Gordon, Kane, & Staiger, 2006).

Like the federal definition of highly qualified, which does not define carefully teacher quality in a way that links it to measureable student outcomes, research provides a plethora of definitions of effective teaching that takes student achievement for granted (Markley, 2004). In fact, researchers have found that defining and describing teacher effectiveness in order to measure it is extremely complex. The problem with research, both in the area of NBC and teacher effectiveness in general, is the narrowness of current operational definitions of effective teaching as well as ambiguity in the use of the terms teacher quality, effective teachers, and teacher effects (Bond et al., 2000; Campbell et al., 2003; Hanushek et al., 1998).

The on-going debate over what an effective teacher is and does makes measuring teacher effectiveness elusive, as there is no generally accepted method for doing so (Goe et al., 2008). Furthermore, because the term effectiveness is an operational construct, the use of different definitions results in obtaining varying process descriptors that shape what needs to be measured (Good, 1979). Consequently, there must be an agreed upon framework for measuring teacher effectiveness. The problem has been that the agreed upon conceptualizations of teacher effectiveness largely have been driven by the available procedures of measurement (Bell et al., 2009; Campbell et al., 2003). As can be seen in the preceding presentation of the historical background of research in the area of effective teaching, when personality measures were available they were used to measure teaching effectiveness, and when measures of teaching styles became available, they became the measure of teaching effectiveness. More recently, fixed effects models have allowed the reemergence of teacher characteristics as one measure of teaching effectiveness. For example, Markley (2004) in his review of research seeking to answer the question "What is an effective teacher?" found that most studies in the early 21st century based their definitions on teacher behaviors. Actions used in defining effective teaching included increasing academic questions, providing instruction to students with different abilities, and promoting higher order thinking skills. Now studies using the value-added model seek to assess the cumulative effects of these teacher actions by applying a mixed-effects approach to analyzing longitudinal standardized test score data across subject areas. The model estimates the effects of schools and individual teachers on student academic achievement (Kupermintz, Shepard, & Linn, 2001; Sanders & Rivers, 1996).

Teacher quality is used synonymously with teacher effectiveness and, as with research on the latter, narrow definitions hamper the study of teacher quality. Furthermore, just as it was found that research on effective teaching followed available measurement procedures, it has been demonstrated that teacher quality is defined by researchers according to what they want to study (Bell et al., 2009; Campbell et al., 2003; Kennedy, 2006; McColskey et al., 2005). One researcher may use the term to refer to credentials and another may use it to delineate academic ability. One solution was to operationalize effective teaching, as Ding and Sherman (2006) did, as one of several variables that influence student achievement. In their model, the variable teacher effectiveness is defined as effective in navigating the interactive student learning process in which strident characteristics as well as teacher efforts, practices, and strategies influence student achievement.

Unfortunately, researchers using teacher quality as the variable for investigation have found the term to be ubiquitous and lacking in specific meaning. For instance, Goe (2007) defined effective teaching as both a component (meaning behavior) and an outcome (meaning raising test scores) of quality teaching. Similarly, Hanushek (2004) defined teacher quality as a distribution between good teachers who produce student gains and bad teachers who do not produce such gains. Hanushek used the definition to illustrate the difference between a good teacher and an average teacher as increasing a student's achievement by 7 percentile rankings in one year. This is the same causal interpretation that value-added models use to describe effective teaching (Kupermintz et al., 2003). Ding and Sherman (2006), whose theoretical framework was presented in chapter I, and Kupermintz (2003) viewed these definitions as circular because teacher quality (effectiveness) is defined and measured by the magnitude of student gains. Likewise, Fenstermacher and Richardson (2005) defined quality teaching as teaching that accords high standards of subject-matter content and methods of practice and successful teaching as teaching that yields the intended learning. They further defined good teaching as teaching that is consistent with morally defensible and rationally sound instructional practices. Again, there is ambiguity between teacher factors and teacher effectiveness that makes it difficult to measure them.

More recently researchers have sought to delineate clearly the concepts of teacher factors (teacher quality) and teacher effectiveness (Blanton et al., 2003; Ding & Sherman, 2006; Odden, Borman, & Fermanich, 2004). The multilevel analytic perspective of these definitions bridges the narrow definitions of teacher variables and teaching results by considering the variety of factors that influence student achievement. As in the past, current advancements in using fixed effects and nested models drive the definition used in these educational models. The multilevel analytical and nested models developed by Odden et al. (2004) and Ding and Sherman (2006) include school, teacher, and student variables to demonstrate the dynamic nature of the learning process. Ding and Sherman's model articulates the difference between teacher factors and teacher effectiveness. In their model, teacher factors are defined as teacher characteristics such as college degrees and years of experience. Conversely, teacher effectiveness is defined by teacher behaviors like working with individual students and instructional practices that increase student achievement. Beyond these two variables, the multilevel dynamic educational model includes school variables (resources, professional, development, instructional leadership, professional community, and cultural climate) and student variables (socioeconomic status, motivation, engagement, and achievement measures) in order to account for the fact that student outcomes result from a set of factors that include more than teacher practices and characteristics (Bell et al., 2009).

National Board for Professional Teaching Definition

As can be seen from the literature review on the history and definition of teacher effectiveness, identifying effective teachers hinges on how it is defined and measured. The National Board for Professional Teaching Standards (NBPTS) used the term highly accomplished teaching synonymously with effective teaching (McClosky et al., 2005; NBPTS, 2001). According to NBPTS, the NBC is a voluntary certification process that measures teacher quality and identifies the most effective teachers (Hakel, Koenig, & Elliott, 2008). The NBPTS defined teacher quality as enhancing student learning and demonstrating the high level of knowledge, skills, abilities, and commitments reflected in 5 core propositions. These propositions are a conceptual framework mirrored by Bell, Little, Croft, and Gitmomer (2009) in the definition they used to measure teaching practice and by Goe et al. (2008) in their 5-point definition of effective teachers. For more than a decade, NBPTS has used the five core propositions to define accomplished teaching to include how all teachers at all grade levels and in all teaching domains demonstrate knowing students well, possessing strong content and pedagogical knowledge, managing a classroom effectively, reflecting deeply on teaching and learning, and engaging in the professional teaching community. Bell et al. (2009), NBPTS (2005), and Goe et al. (2008) attempted to bridge the narrow definitions that focus purely on processes and products.

What is left out of the above definitions and measurement of effective teaching is student achievement outcome data. Both groups have taken for granted that effective teaching will automatically yield positive student outcomes. Even with NBC quickly becoming identified by educational leaders after its inception as providing effective methods, practices, and processes for building knowledge from practice, those leaders' beliefs were not substantiated by empirical studies. It has only been in the last decade that researchers have returned to quantitative methods to assess the effect of NBC on student outcomes (achievement data). As outlined in chapter I, Bond et al. (2000), after an extensive review of the literature on teacher expertise, developed a conceptual model of teaching expertise based on Schulman's (1986) seminal work regarding pedagogical content knowledge (PCK). PCK involves several features relevant to the study of teacher effects on student achievement, the teacher's knowledge of what is being taught, knowledge of instruction, knowledge of the skills, gaps, and preconceptions students may bring to the subject, and knowledge of the diverse instructional strategies needed to teach for understanding (Rowan, Correnti, & Miller, 2002).

In developing a construct of teacher expertise, Bond et al. (2000) clarified 15 dimensions of accomplished teaching from over 20 years worth of study on effective teaching that fell into three broad areas that were the focus of their study: Comparative Teaching Practices, Comparative Teaching Outcomes, and Comparative Professional Activities. Bond et al. then developed protocols to evaluate board certified and nonboard certified teachers on these dimensions. Using (a) reviews of teacher assignments and student work, (b) student interviews and questionnaires, and (c) classroom observations that evaluated student self-efficacy and motivation as well as classroom climate and environment, the researchers found that board certified teachers performed higher in all 15 dimensions (Table 1) than candidates who unsuccessfully attempted certification and were statistically significantly higher in 13 of the dimensions of teacher expertise. Student motivation and self-efficacy and a passion for teaching only were negligibly higher for NBCTs than unsuccessful candidates (see asterisked items in Table 1).

Although the conceptual framework of Bond and his colleagues lends itself to the qualitative studies of NBC, where observations, student samples, and surveys are used to look at teacher effectiveness, it does not provide the framework for a connection between teacher factors (effects) and effective teaching as defined by Ding and Sherman (2006). For this reason, unlike past research, the Ding and Sherman conceptual framework is used in this meta-analysis that investigated teacher factors (effects) as the variables that influence student achievement or as Ding and Sherman's model defines it teacher effectiveness.

Table 1
Dimensions of Teaching Expertise

Dimension	Characteristics
Use of Knowledge	Expert teachers' knowledge is more integrated; experts connect new information with their own prior knowledge and that of their students.
Deep Representations	Expert teachers recognize patterns of student responses and use the patterns to interpret events and plan instruction.
Improvisation	Expert teachers define and redefine instructional and curricular issues to reach resourceful and insightful solutions that may not occur to others.
Problem Solving	Expert teachers adapt and modify their instruction during the flow of the lesson.
Challenge of Objectives	Expert teachers set demanding goals on the basis of students' current competencies.
Decisions	Expert teachers understand the hows and whys of student success and make decisions based on this understanding.
Classroom Climate	Expert teachers create optimal environments for learning by interpreting and using students' verbal and nonverbal behavior to anticipate and prevent disruption using management procedures they developed to accomplish their goals.
Multidimensional Perception	Expert teachers filter relevant information from less relevant information and use it to address their instructional goals and encourage academic engagement.
Sensitivity to Context	Expert teachers are sensitive to task demands and social situations in the classroom and use that knowledge to guide instruction and management decisions.
Monitoring Learning and Providing feedback	Expert teachers diagnose students' interpretations and tailor feedback to correct students' misunderstandings or to help them create new learning connections.
Test Hypothesis	Expert teachers evaluate the information they gather related to potential actions, reactions and adjustments to lessons.
Respect for Students	Expert teachers respect and care for students as learners and as people, recognizing barriers to learning and simultaneously empowering them to overcome barriers.
Passion for Teaching and Learning*	Expert teachers display passion that is closely linked to their commitment to student success and their sense of responsibility and love of the subjects they teach.
Motivation and Self-efficacy*	Expert teachers encourage content mastery motivated by students' personal sense of intrinsic satisfaction.
Outcome of Lessons: Surface and Deep	Expert teachers encourage all students to achieve and understand content at increasingly complex levels.
Outcome of Lessons: Achievement	Expert teachers' effect should be evident in the depth and quality of student responses elicited by teaching encounters.

Note. Based upon the literature on teaching expertise from Bond, et al. (2000). *The certification system of the National Board for Professional Teaching Standards: A Construct and consequential validity study.* University of North Carolina at Greensboro: Center for Educational Research and Evaluation.

What is clear from the review of the literature in this section is that effective teaching is multidimensional. It must be defined as encompassing both meeting the expectations of the teacher's role (attributes and practices) and the results of teacher actions on student achievement (Blanton, Sindelar, & Correa, 2006).

Models and Measures

Previous as well as current research agendas have focused on accountability and performance standards that cover both teacher factors, such as experience and advanced degrees, as well as standardized assessments. As with all history, the past has influenced the present, which is why the next subsections focus on methods that have been and are used to investigate teacher effectiveness.

Process-product Research

The process-product research approach, considered to be the first successful empirical method used in the field of teacher effectiveness, was an important source of data concerning teacher factors (Yates, Chandler, & Westwood, 1987). From the 1960s to the early 1980s, this research method provided quantitative analysis of the relationship between teaching skills, or effects, and student achievement. Typically centered on teaching and achievement in reading and mathematics, this type of research used detailed observations of how teachers functioned as the independent variable in a classroom. Observable, discrete teacher behavior was considered an effect rather than a cause of student achievement (Good & Grouws, 1977; Medley, 1977b; Yates et al., 1987) Behavioral categories classified product, process, predictor, and context variables. Student outcomes were the product variables and included both student achievement and attitudes toward learning. Teaching methods served as the process variables that enhanced student learning (product). Student characteristics and prior knowledge were the presage variables that are those variables associated with the teacher. These variables are proposed to affect the behavior of the teacher in the classroom. They are the teacher factors in Ding and Sherman’s model (2006). Context variables were factors that influenced the effects of teaching and student outcomes (Fenstermacher & Richardson, 2005; Gage & Needels, 1989; Seidel & Shavel, 2007).

It is the process-product method that ushered in the use of standardized test scores as a measure of teacher effectiveness. Used as the initial identifier of effective teachers to be observed, achievement scores also were used to evaluate empirically the relationship between teacher behavior and the quantity of student learning (Medley, 1977b). Process-product research has played a key role in the development of the NBPTS assessment process, in the studies that control for context variables, and in research methods that use standardized test scores as a measure of effectiveness.

Lee Shulman (1986a), who was a principal designer of the NBPTS assessment program, gave an overview of the implications of the process-product paradigm in a well-known chapter of the *Handbook of Research on Teaching*. Based in behaviorist psychology, Shulman regarded the process-product method of studying teacher effectiveness as reducing classrooms to discrete events and behaviors that could be observed, counted, and analyzed for the purpose of producing better student learning. This view of the relationship between what teachers do and what children learn was problematic for Shulman who realized that, in an effort to identify effective teachers, the subject matter and other intervening variables gradually were being ignored (Lederman & Niess, 2001; Yates et al., 1987). As a result of believing that research in teaching had become too generic, Shulman formulated a new paradigm that combined six domains of knowledge. Schulman’s (1986b) PCK became the basis for the development of the five core propositions of NBPTS. PCK involves several critical features a teacher must possess in order to affect student achievement. The first feature is teacher knowledge of what is being taught; however, it is not enough to know the subject. There is a second necessary feature: knowledge of instruction. Without being able to translate subject knowledge into effective instruction for students, academic progress will be affected. Ultimately, teachers also need an understanding of another PCK feature, the skills, gaps, and preconceptions students may bring to the subject and the related, diverse instructional strategies needed to teach for understanding (Rowan et al., 2002).

At the same time that Shulman was working with the NBPTS to develop the standards for accomplished teaching he was also developing PCK . It, therefore, is not surprising that one tenant of what the board believed accomplished teachers should know and be able to do is possess pedagogical content knowledge (NBPTS, 2002). It also is not unexpected that researchers began to move away from a focus on teacher behaviors and indirect measures of teachers' subject-matter knowledge, including standardized test scores. PCK research agendas focused on what teachers wanted students to know and be able to do and on the study of the decisions teachers made regarding subject matter goals and content selection and representation choices (Lederman & Neiss, 2001).

This last outcome of the move away from process-product research also is reflected in the National Board's choice to exclude standardized tests as a measure of accomplished teaching. It also influenced the research agenda of Bond et al. (2000), whose validation study has been criticized on the basis of using process-product observation methods to validate the effect of a teacher's adherence to the 5 core propositions on student achievement using student work products and writing samples (Cunningham & Stone, 2005). As mentioned in the previous section, NBPTS' lack of empirical evidence for validity was the impetus for a decade of research intended to validate certification, which was the focus of the meta-analysis.

Value-added Research

Another research model that is associated with NBPTS certification process is value-added research (Clotfelter, Ladd, & Vigdor 2007; Goldhaber & Anthony, 2004; Harris & Sass, 2007; Kane & Staiger, 2008a; Stone, 2002). Value-added models (VAMs) are a relatively new statistical method of estimating the contributions of schools, teachers, or both to student learning as represented by test score trajectories for purposes of accountability. The intention is to make causal inferences by correcting for nonrandom pairings of students with teachers and schools (Ballou, Sanders, & Wright, 2004; Goldhaber & Hansen, 2005; Harris, 2005). Teachers whose students make greater than expected growth have high value-added ranking, which is judged via a scale score that results from the VAM analysis. The number describes the difference between one teacher's performance and a typical teacher's performance with respect to the average growth of their students on standardized tests (Braun, 2005). Randomization is needed for this method to be equitable so that each teacher has an equal chance of having a mix of student abilities in his or her class. The fact that randomization is not feasible in most districts means that contextual variables like socioeconomic status and demographics need to be controlled.

In order to tackle the problem of nonrandom assignment of students to teachers and teachers to schools, value-added modeling adjusts for preexisting differences among students, using a student's history of test performance as a substitute for omitted background variables. In experimental terms, each student and teacher serves as her or his own group (Ballou et al., 2004; Braun, 2005; Harris, 2008). This method of using student’s prior achievement as a proxy for family and neighborhood variables is considered a blocking factor that enables the VAM to estimate the effects of teachers, schools, and school systems (Kupermintz et al., 2001). There is concern, however, that blocking may mask reasons for student gains other than teacher quality. For instance, the value-added scores of teachers who consistently are assigned high-achieving students may be upwardly biased (Amrein-Beardsley, 2008; Ballou, 2008).

Contextual factors are another concern because student learning is not just a function of a teacher’s effectiveness or a student’s ability and effort. Overall classroom climate and peer-to-peer factors, such as classroom disruptions and the positive influence high-achieving students have on their peers, are captured under the category titled teacher factors in the value-added model. These and other time-varying components, such as administrative and support staff, and neighborhood and community factors, diminish the useful of the value-added approach as a means of analyzing teacher effectiveness because it is unclear how well using value-added methods controls for these cofounding variables (Ballou, 2005a; Kupermintz et al., 2001).

Missing data also are an issue for most VAMs because they require complete, high-quality longitudinal data that frequently are unavailable. When student data are missing, it is functionally impossible to measure learning gains, even though the claims of Sanders et al. (2002) that the system can operate in the absence of data. Furthermore, missing or faulty data can have a negative effect on the precision and stability of value-added estimates and also can contribute to bias (Amrein-Beardsley, 2008). This is particularly important in districts with high mobility where many students have an insufficient history of prior achievement to be included in a value-added analysis. As with contextual factors, such as socioeconomic characteristics of the classroom including poverty and race, most researchers employ a variety of fixed effects along with longitudinal data to reduce the potential for omitted variables to bias estimates of teacher effectiveness. Adding these contextual effects into the model, however, changes the function of the value-added model because it restricts the inferences that can be drawn about the effectiveness of different teachers (Amrein-Beardsley, 2008; Ballou, 2005b). To address this concern of inadequate data, a teacher is assumed to perform at her or his school system average, which can lead to false positive and false negative classifications. Additionally, unaccounted for data may potentially bias the results because they ignore students who tend to be underperforming or who do not participate in large-scale assessments (Amrein-Beardsley, 2008; Kupermintz et al., 2001).

Although research using value-added measures provides away empirically to evaluate a teacher's contribution to student learning and demonstrate that there is a wide variance in what teachers contribute to students' gain scores, it gives no indication of how their instruction contributed to or impeded student learning (Goe et al., 2008; Harris, 2008). Additionally, little is known about the validity of their scores for identifying effective teaching (Amrein-Beardsley, 2009; Goe et al., 2008). Validity for VAMs depends on how accurately the model captures an individual teacher's contribution to student achievement through growth on standardized achievement tests. What currently is needed to assess whether VAMs do indeed capture teacher contributions is an active research program that focuses on contextual and background variables that pose a threat to the validity and the misclassification of teachers (Amrein-Beardsley, 2008; Goe et al., 2008; Kupermintz, 2003).

An additional validity concern is the lack of peer review of VAMs. Researchers have found it impossible to acquire the necessary computational algorithms for the most publicized VAM, the Educational Value Added Accountability System (EVAAS), formerly the Tennessee Value Added Assessment System (TAAS), because the developer holds them as proprietary information (Amrein-Beardsley, 2008; Kupermintz et al, 2001). Peer review commonly is consider a key component of research validation as is the replication of research that is accepted as an essential part of scientific research that ensure the validity and reliability of research methods. By ignoring this valuable scientific standard, EVAAS developers have ensured that the sale of their system will not be hampered by any research that may call into question its validity and in so doing are exposing thousands of teachers to accountability measures that may misidentify them as ineffective (Goe et al., 2008).

A final concern is that individual teacher value-added score is unstable over time. (Ballou, 2005b; Kodel & Betts, 2007). Although instability may be due to genuine changes in teacher effectiveness, some portion of it may be the result of measurement error. Variability is still apparent once test gain measurement errors are corrected with the degree of stability ranging from 50% to 90% as measured by teachers staying in the same performance level. Stability also requires that students and school factors be omitted from comparisons of teacher factors (McCaffey, Lockwood, & Sass, 2008); however, as presented earlier, such omissions restrict the inferences that can be drawn about the effectiveness of different teachers (Amrein-Beardsley, 2008; Ballou, 2005b). Imprecision of value-added models that only make it possible to distinguish between the very highest and very lowest level of teacher effectiveness also may play an important role in the instability of teacher value-added scores over time (McCaffey, et al., 2008).

Amrein-Beardsley (2008) concluded her study by linking various value-added research that put NBC to the test. Her reanalysis of the Sanders, Ashton, and Wright's (2005) study, comparing 4 years of test scores of NBCTs and nonNBCTs on mathematics and reading outcomes, reversed their findings. Using more traditional methods of analysis, Amrein-Beardsley confirmed that students of NBCTs learn statistically significantly more than the students of nonNBCTs and that the value-added method both reduced the number of statistical findings and weakened effect sizes. Other researchers have come to similar conclusion (Hakel et al., 2008; Vandervoor et al., 2004) making it appropriate to look at methodology as a moderator variable. Unfortunately, descriptive statistics are not necessary for the value-added calculations and were not reported in sufficient detail to allow VAMs used to study the effect of NBC on student achievement to be included in the meta-analysis.

Synopsis of Standardized Tests as a Measure of Teacher Effectiveness

The presentation of standardized test scores in this section provides the rationale for their use as the dependent variable in this meta-analysis and the studies included in it. Standardized tests measure a standardized set of broadly based educational outcomes, uses standard directions, along with standard scoring procedures. These tests are standardized in order to provide a comparison of student scores with those of similar students who have taken the test under the same conditions (Gronlund, 2006).

As discussed in the earlier section titled *Process-product Research,* NBPTS came of age during a time when researchers were moving away from using indirect measures of teachers' subject matter knowledge including standardized test scores (Cumming & Maxwell, 1999). As a result of that change and a shift in educational policy toward using authentic assessment, the decision was made that the NBC process would not include students' standardized test scores (Hagel et al., 2008; Ingvarson & Hattie, 2008). Again, this decision came under fire in the early 2000s because there was little empirical evidence that the NBC process was more effective than the more traditional licensing exams in identifying effective teachers (Predrosky, 2001; Stone, 2002). In fact, at that time, the only study to attempt to validate the NBC process was by Bond et al. (2000), presented earlier in the section titled *National Board for Professional Teaching.* Like the NBPTS, these researchers argued that standardized tests in state accountability plans are narrow, inadequate, and arbitrary, which ended up leading critics to challenge the Board to put their certification method to the test (Predrosky, 2001; Stone, 2002). In response to criticism, the Board commissioned a number of studies to assess the effect of certification on student achievement. In keeping with standards-based accountability of the *No Child Left Behind Act of 2001*, the majority of studies related to NBC and student outcomes have focused on achievement scores as an identifier of effective teaching.

Notwithstanding the ever growing body of research that analyzes NBC and its relationship to student achievement (Vandevoort et al., 2004), there have been no conclusive findings (Clotfelter, Ladd, & Vigdor, 2006; Harris & Sass, 2007; Sanders et al., 2005; Vandevoort et al., 2004). These varied findings mirror three decades of research that suggests considerable heterogeneity in teacher effectiveness on student achievement. One main reason for the inconsistent findings is that there are issues regarding the interpretation of findings. As provided in chapter I, when reviewing studies that examined the relationship between board certification and student achievement, the National Research Council (Hakel et al., 2008) found that only one study (Cantrell et al., 2007) of the seven they reviewed did not have interpretation issues. One important method of measuring this variation of teacher effectiveness on student outcomes is identifying teacher, student, and school variables that can account for the variance. These variables, called fixed effects, are presented in the next section titled *Various Influences on Study Outcomes.* There are, however, several other important considerations in using student achievement as a criterion for effective teaching that can influence the outcome of both the primary studies in this meta-analysis and the meta-analysis itself.

The simplest and most frequently employed approach to estimating teacher effectiveness is the pretest-posttest design used to measure the degree of change occurring as a result of instruction (Knapp & Schafer, 2009; Kupermintz et al., 2001; Popham, 1999). One concern with this method also was an issue in the earlier subsection entitled *Value-added Research.* In order to isolate properly and nullify confounding variables, the students must be assigned randomly to teachers, which are not feasible in most schools and school districts (Braun, 2005; Dimitrov & Rumrill, 2003; Goe et al., 2008). If random assignment were possible, there would be more confidence in the resulting use of test scores to assign an effective or ineffective label to a teacher; however, there would still be a problem with the possibility that using gains from one year to the next relies on the assumption that assignments are uncorrelated with previous instruction (Kupermintz et al., 2001). Families choose communities and schools to give their students the best advantage. Principals place teachers with students they believe will benefit from or receive the least harm from their instruction. There is differential and sometimes preferential means by which districts assign teachers to schools. All of these different factors regarding developing assignment of students to classrooms have the potential of confounding student achievement and teacher effectiveness (Braun, 2005; Kane & Staiger, 2008b).

Shavelson, Webb, and Burstein (1986) listed several other factors that obstruct the study of teacher effectiveness using pretest-posttest scores. First, curriculum and standardized tests rarely are aligned resulting in instruction that adheres to the test's curriculum. Added to this concern is the problem that no test or assessment is likely to cover the full domain of content standards. Even those aspects that are covered will vary in degree and depth of coverage. For this reason, generalizing from the content of the test to the domain of the standards requires an adequate evaluation of alignment that makes it clear which aspects of the content standards are left uncovered by the test, which are covered only lightly, and which receive the greatest emphasis (Kupermintz et al., 2001). This alignment especially is important because if tests do not cover what teachers actually teach, the most sophisticated analysis will not yield credible estimates of teacher effectiveness (Ballou, 2008).

Shavelson et al.'s (1986) second concern is interconnected with Goe et al.'s (2008) concern that a single score measuring a teacher's effectiveness assumes that all or nearly all a student learns in a year is the product of a single teacher's effort. Likewise, Shavelson et al. believed that aggregating student scores across all students in a classroom ignored differential learning among students, masking a teacher's contribution to student learning. They also believed that scores are used unfairly to judge a teacher's effectiveness on learning because equating performance on a test with a students' knowledge of the subject overlooked the influence of test-taking strategies, motivation, and attitude toward testing. Added to this problem is the concern that when gain scores are equated with teacher effectiveness, it becomes impossible to differentiate between instructional practices that promote learning and those that narrowly teach to the test (Kupermintz et al., 2001).

One final concern for Shavelson, et al. (1986) was that standardized tests are strictly summative and are not adequate representations of student's cognitive growth. This issue of assessments reflecting cognitive growth also is reflected in the results of the type of standardized test used. For instance, norm-referenced tests are not aligned to state standards, which make inferences about the effectiveness of districts, schools, and teachers based on such tests questionable. Grade-level criterion-referenced tests are linked to state standards but are not sensitive measures of cognitive growth. In fact, the more academically advanced a student is, the less apparent is their progress on criterion-referenced test (Amrein-Beardsley, 2008).

Although tests can be scored objectively, tests are incomplete measures of student achievement making the resulting inferences that can be drawn from them subjective. Avoiding the potential biases of estimating teacher effects based on students' scores requires rigorous experimental research design procedures that account for variables beyond their control (Kane & Staiger, 2008a). Consequently, the following section provides a detailed overview of the factors that account for student score variability that were coded as moderator variables in this study. The need for rigorous experimental research design procedures also was the basis for coding and analyzing the quality of studies included in the meta-analysis.

Various Influences on Study Outcomes

This section provides an overview of research on the relationship between the categories of factors that the research literature (Ding & Sherman, 2006; Odden et al., 2004) labels teacher factors and student factors and their influence on student outcomes and methodological quality. These factors provided the basis for the moderator variables that were used to develop the coding categories in the current meta-analysis. More specifically, the first subsection focuses on instructional methods, teaching behaviors, and indicators of teacher quality, whereas the next subsection focuses on the variables of race and socioeconomic status. Finally, the last subsection pertains to the variable of study quality.

The third research question of the meta-analysis explored the extent to which study features moderate the relationship between certification status and academic achievement. There is a large body of research demonstrating variance in teacher effectiveness that can only be accounted for by unmeasured teacher, student, and context heterogeneity. As Hanushek et al. (1998) pointed out, current research indicates that teacher heterogeneity is the most important component of achievement variation. Because it is not possible to review all existing literature on factors influencing achievement variation, only those that serve as potential moderator variables in the meta-analysis are reviewed in the subsequent subsections of this chapter. Variables that research has shown to play a role in the variance in results of study effectiveness will be provided but details will not be given.

Teacher Characteristics Influencing Student Achievement

Earlier in this chapter, there was a focus on the ambiguity between the terms teacher effectiveness and teacher effect. The current meta-analysis used the conceptual framework of Ding and Sherman (2006) to define the category teacher factors as teacher characteristics such as college degrees and years of experience and teacher effectiveness as teacher behaviors like working with individual students and instructional practices that produce growth in student achievement. Including teacher factors in studies of teacher effectiveness is important because without them the effect-size estimates of student differences in achievement may be biased (Wayne & Young, 2003).

Although research has shown that teachers considerably influence student achievement, studies indicate that there is substantial variation in teacher characteristics and practice. Research also demonstrates that this variation in teacher effectiveness remains largely unexplained by common measures of teacher characteristics (Clotfleter et al., 2007; Harris & Sass, 2009; Munoz & Chang, 2007). The aspects of teacher characteristics that have been found to bear some relationship to student achievement across studies include licensure and measures used for licensure, such as years of experience, advance degrees, verbal ability, and personal traits such as gender and race. All of these factors are reviewed here and, except for verbal ability, which did not appear in any of the studies used in this meta-analysis, were coded for this study.

With the NCLB requirement that every student receive instruction from a “highly qualified” teacher, current research has given increased attention to the relationship between teacher licensure (state certification), examination scores, and tests of verbal skills that have been used to grant teacher licenses (Wayne & Youngs, 2003). One study conducted by Huang and Moon (2009) found that licensure status was not statistically significant with regard to student achievement. Other research conducted by Wayne and

Youngs on subject certifications given by states had opposite results finding that subject-specific certifications matter.

In their synthesis of the literature on these verbal ability and licensing tests, Wayne and Youngs (2003) stated that joint interpretation of studies that assessed the importance of teacher test scores and scores on verbal tests indicated that students learn more from teachers with higher test scores. Their findings were reflected in review of literature conducted by Darling-Hammond (2000) but refuted by Goldhaber and Hansen (2008) who found that licensure exams do not function as a good screen for teacher effectiveness. Likewise, a synthesis of the literature by Aloe and Becker (2009) did not support this conclusion in regard to verbal ability whereas Darling-Hammond and Youngs (2002) did support the previous findings of verbal ability being associated with increased student achievement. Darling-Hammond and Youngs' review of correlation and regression studies indicated that verbal ability is not strongly correlated with student outcomes. One finding of concern was that Goldhaber and Hansen (2009) observed that a disproportionate number of African American candidates are among those who fail licensure tests. They further found that these tests function differently among African American and male teachers as measured by student outcomes, with African American and male groups performing higher on some portions of the test and lower on others than European American and female teachers.

Wayne and Youngs (2003) also reviewed the literature on the effect of degrees and course work on teacher effectiveness and concluded that high school students learn more from teachers with course work and degrees in mathematics. Although Goldhaber and Anthony (2003) found that degrees in subjects different from those being taught by a teacher had little effect on student outcomes. Using a value-added model to correlate attainment of degrees with teacher effectiveness, Harris and Sass’ (2009) findings were consistent with those of the previous two research syntheses, suggesting that other factors may play a more prominent role in determining student achievement. Clotfelter et al. (2007), using fixed effects models, found that teachers with advance degrees were less effective at raising student test scores, as did Darling-Hammond (2000).

The findings of Clotfelter et al. (2006, 2007) on teacher experience indicate that more experienced teachers are more effective, with the greatest gains in becoming effective occurring in the first few years. These results are supported by the reviews of literature completed by Goldhaber and Anthony (2003) and Darling-Hammond (2000) who found that studies allowing for the nonlinear relationship between student achievement and teacher experience provide convincing evidence that the value of teacher experience matters mostly during the first 5 years in the classroom when student gains increase with each successive year a person teaches. After that time student gains for a teacher do not continue to increase or decline with more years of experience. Harris (2008) conducted a study of the value-added methodology, and he went as far as to say that teacher experience is the characteristic that most clearly is related to teacher effectiveness. The findings of the researchers cited earlier in this paragraph were not supported by the value-added studies of Munoz and Chang (2007) and Huang and Moon (2009), who found that years of experience had little relationship with student achievement. Huang and Moon did find, however, that years of experience in the same grade level had a positive relationship with student reading scores. Hanushek et al. (1998) also found teacher experience to only be a small component of variations in teacher effectiveness.

Clotfelter et al. (2007) explored the relationship between student outcomes and the gender and race of a teacher. They found that women tend to be more effective than men and African American teachers are less effective than their European American counterparts. When race is a shared characteristic between student and teacher there are more positive outcomes in both reading and mathematics. Goldhaber and Hansen (2009) also found that African American teachers have a greater relationship with student achievement of minority students than European American teachers even when the African American teacher has a low score on licensure test. The value-added study of Munoz and Chang (2007) refuted this finding because they found no association between race and student achievement.

Some of the research models in the primary studies in this meta-analysis use only students' prior tests scores to calculate teacher effectiveness, whereas other models included teacher gender and race, and still others include teacher experience, degrees, and licensure status. As presented in the subsection *Value-added Research*, omitting these variables restrict the inferences that can be drawn about the effectiveness of different teachers (Amrein-Beardsley, 2008; Ballou, 2005b). Therefore, those variables were coded in this meta-analysis both to assess their relationship with certification status and as a measure of study quality.

Student Variables Influencing Student Achievement

Ding and Sherman's (2006) conceptual framework goes beyond the variables of teacher factors and teacher effectiveness. The multilevel dynamic educational model includes context and student variables in order to account for the fact that student outcomes result from a number of variables beyond teacher practices and characteristics. Hattie (2003), after an extensive review of the literature conducted for a research synthesis, used effect sizes to estimate the relationship between these variables and student achievement. He calculated that student characteristics account for 50% of the variance in achievement and that, except for 30% accounted for by the teacher, the remainder (20%) came from contextual factors. This subsection contains a review of the literature on these two groups of variables.

One of the most frequently used statistical controls employed to enhance inferences is socioeconomic status (SES). A myriad of research has documented the relationship between SES and student achievement, not the least of which is the Coleman Report (Coleman, 1966) that erroneously indicated that SES was a stronger determinant of academic achievement than an effective teacher (Benigno, 2005). Free-and-reduced lunch (FRL) is often used to statistically control for SES on student achievement in order to increase statistical power and enhance arguments of causation (Harwell & LeBeau, 2010). Clotfelter et al. (2007), using FRL status and parents who are only high school graduates as a proxy for SES, found that larger concentrations of poor students in a teacher's classroom decreases achievement scores. Similarly, Stewart (2008) and Hanushek, Kain, Markman, and Rivkin (2002) showed that SES statistically significantly was associated with academic achievement. One final study of SES found that the relationship of SES with student achievement remains stable during the elementary years but increases rapidly up to the tenth grade (Caro, 2009).

According to Noguera (2008), the variable of race continues to be a factor in student achievement notwithstanding the NCLB mandate to devise means to ensure that student achievement increases regardless of background. Hedwig (2007) supported this statement with a study using ordinary least squares (OLS) and hierarchical linear modeling (HLM) in which he concluded that the racial makeup of a school has an important relationship with student achievement. Stewart (2007), when studying the influence of school-level and individual-level factors on academic achievement, also found race to be correlated with student achievement scores.

Inconsistent research results frequently are reported regarding the relationship between gender and student achievement. For instance, gender was not statistically significantly related to academic achievement in Stewart's (2007) study, but statistically significant differences were found between boys and girls in a study of the influence of gender, academic achievement, and nonschool factors upon pre-adolescent self-concept by Hay, Ashman, and Kraayenoord (1998). Watson, Kehler, and Martino (2010), studying teacher characteristics and student achievement gains, asserted that it is an established fact that boys perform less well than girls on literacy benchmark or standardized tests. Both the National Assessment of Education Progress (2009) and Francis and Skelton (2005) agreed that, on average, girls outperform boys on achievement tests with largest gap in reading.

Student demographics, including socioeconomic status (SES), gender, and race, have all been researched thoroughly in regard to their association with student achievement. For this reason, they were coded as moderator variables in this meta-analysis.

Contextual Factors Influencing Student Achievement

The influence of economic status and other background variables are deemed by some in the research community as necessary for the measurement of teacher effectiveness (Ballou et al., 2005; Hanushek et al., 1998). Because studies have shown that school and classroom context are related to educational outcomes, these variables are presented in this subsection and were coded in the current meta-analysis. An overview of the primary studies in this meta-analysis provided the general categories of urban, suburban, metropolitan, and rural schools and neighborhoods. FRL was the common proxy in the studies for SES when assessing school and neighborhood contexts.

Although there is limited empirical research on how highly effective teachers perform in different settings, what there is points to the assertion that teacher quality is context specific (Goldhaber & Anthony, 2003). One content-specific classroom characteristic that has been researched extensively since 2000 is class size, with optimal classes being defined as having less than 20 students. The results of the research, however, are varied. For instance, Hanushek et al.’s (1998) study of teachers and schools and their relationship with academic achievement demonstrated that class size is related to reading and mathematics achievement of students from low-income families, but the relationship declines with increases in grade level. Clotfelter et al. (2007) found that reducing class size by 5 students statistically significantly increases student achievement, whereas Darling-Hammond (2000) found weak and insignificant increases, when studying teacher quality and student achievement. Odden et al. (2004), in their review of the literature, confirmed that the relationship of class size persists into later grades and that minority students receive important benefits from placement in smaller classes.

A number of studies have examined the relationship of peers and school factors with student achievement. For instance, Stewart (2007) found positive peer associations are connected with increased student achievement, and Hanushek et al. (2002) demonstrated that the relationship was statistically significant across the test score distribution. In a review of the literature, Jargowski and El Komi (2009) also showed that the higher the mean test score of classmates, the higher the achievement level of the student. This relationship was nonlinear and decreased as the mean of peers rose. Additionally, their findings indicated that struggling learners were affected negatively by underperforming classmates. For high achievers, Hanushek et al.'s (2002) results show that they were unaffected by variation in the percentage of top achievers. This finding is supported by the research of Burke and Sass (2008), when studying classroom peer factors and student achievement. Card and Rothstein (2007) also demonstrated, with their study of racial segregation and the African American-European American test score gap, that student achievement depends on the expectations and achievement of peers, demonstrating that segregation has a negative relationship with student achievement. Stewart's (2007) research indicates that the factors explored by Card and Rothstein can be mitigated by cohesive, inviting school environments.

Neighborhoods, as well as schools, are contextual variables that have been researched in terms of their effect on student achievement (Jargowski & El Komi, 2009). Neighborhoods are not static, in fact 30% of the nations poorest children have attended three schools by the third grade and frequent mobility has been demonstrated to be a factor in poor student achievement (Berliner, 2009). Card and Rothstein (2007), when studying racial segregation and the African American-European American test score gap, found robust evidence that race is also a contextual variable in more segregated cities because the African American-European American test gap is larger there. Whereas Card and Rothstein’s research provided evidence that neighborhoods played a greater role than schools in student achievement in high poverty and minority cities, Jargowski and El Komi (2009), in their study of school context and neighborhood factors on student achievement, found quite the opposite.

Research regarding family influences on academic achievement is mixed. As a result of researching the influence of teacher differences on academic achievement, Hattie (2003) reckoned that family accounts for 5% to 10% of the variance in student outcomes. Family characteristics including education and income are a strong predictor of student test scores (Card & Rothstein, 2007). Clotfelter et al. (2007) found that parental education levels exert a larger influence in reading than in mathematics.

As with teacher and student variables, the consequences of omitting contextual variables may cause discrepant substantial upward bias in the magnitude of teacher effects (Palardy, 2010). Therefore, every effort needs to be made, in comparisons of teacher effectiveness, to avoid erroneous results by including all relevant variables in the initial model. For this reason, in addition to estimating the relationship between certification status and student achievement, the context variables presented in this section were coded for all primary studies selected for this meta-analysis.

Methodological Quality

In their book, *Summing Up: The Science of Reviewing Research,* Light and Pillemer (1984) presented research design as a source of variance. They pointed out that results can be modest, negligible, or large depending on the research design. Light and Pillemer also indicated that, like the other moderator variables presented in this section, there are contradictions in the research. Their point is that numerous studies have found a clear relationship between research design and results, making it important to examine the relationship. The following addresses the features of studies that researchers have concluded influence findings and that were coded in this study in order to investigate the relationship between methodological variations and study outcomes (Lipsey & Wilson, 2001).

Studying design quality relationships with study results can be done posteriori by coding for design aspects of each study and then demonstrating whether or not the outcomes of studies are related to how the research was conducted (Cooper, 1998). Consequently, the decision was made a priori for this meta-analysis to code (a) the statistical methodology used in the study, (b) if the study used pretest-posttest design, and (c) the type of assignment to a research group. By carefully enumerating the study descriptors to be used to assess design quality, empirical comparisons of differences can be used to assess how well a study investigated the relationship between certification status and student achievement (Rosenthal, 1991). The above choice of descriptors was based on a broad overview of a sample of the studies to be coded in order to investigate what information was recorded frequently enough to justify the coding effort (Lipsey & Wilson, 2001).

Methodological concerns regarding value-added studies that were presented earlier under the subject heading *Value-added Research* will not be given here. Another concern regarding VAMs pointed out by Light and Pillemer (1984) is that omitting controls is one way to improve results. This problem, however, is not unique to VAMs. For example, Palardy (2010) concluded from his research on a multilinear cross-random-effects growth model for estimating teacher and school factors that even a small degree of unmodeled nonlinearity can result in a significant upward bias in the magnitude of the teacher effect. In order to avoid this problem of omitted variable bias, methodologists with student-level panel data often employ specifications that include a variety of fixed effects to control for school, peer, and background inputs (Ballou, 2008; Harris & Sass, 2007). Because fixed effects have been shown to yield consistent estimates, methodologies that used them, such as HLM and regression models, are considered high quality for the purposes of this meta-analysis. Studies that do not contain fixed effects, including independent samples *t* tests, OLS studies are considered to be of lower quality.

Random samples, which are samples chosen from a given population in a way that makes sure that every person has an equal and independent chance of being selected for the sample (Weinberg & Abramowitz, 2002), are the best means of ensuring that effects attribute to teachers are not misidentified. As mentioned earlier, randomization is not feasible in most school settings for a variety of reasons. For this reason, two other methods are used frequently to control for biased estimations that may result from nonrandom sampling. The first is matching, which approximates randomization as closely as possible. Matching methods generally are used to select well-matched groups from both the control and experimental samples in order to reduce bias due to covariates in the context of causal inferences (Stuart & Rubin, 2008). Eight primary studies in this meta-analysis, presented previously in this chapter, matched NBCTs and nonNBCTs on teacher factors, including gender, years of experience, and advance degrees.

The second method, blocked sampling, involves dividing the group of experimental units into homogeneous blocks of equal size and then assigning them to a treatment group. This method controls for any preexisting differences between the two groups that make them unbalanced. Interpreting the results from an unbalanced trial may lead to reaching biased conclusions about teacher effectiveness (Ariel & Farrington, 2010). One study included in this meta-analysis used the blocked samples method to assign randomly students of NBCTs and nonNBCTs. The remaining six used nonrandom assignments of students and frequently were unbalanced with considerably larger numbers of nonNBCTs. The first two methods control for bias and, therefore, are considered more rigorous for the purpose of determining study quality in this meta-analysis.

Notwithstanding research that raises concerns about pretest-posttest designs, they are the preferred method of comparing two experimental groups when measuring the degree of change occurring as the result of a treatment. The primary studies acquired for this meta-analysis that employed this design used gain scores (difference between the posttest mean and the pretest mean) to compare the students of nonNBCTs to the students of NBCTs in order to analyse differences. Traditional statistical methods that are used to compare groups with pretest and posttest data are the analysis of variance (ANOVA) and analysis of covariance (ANCOVA). These both were utilized by the researchers in the primary studies included with the meta-analysis. Using pretest scores in both these methods helps to reduce error variance, which produces more powerful findings than studies with no pretest data (Dimitrov & Rumrill, 2003). For this reason, the primary studies using the pretest-posttest design were considered higher quality than those that did not when coding study quality in the meta-analysis.

Summary

Teacher effectiveness has been the focus of considerable research. That research frequently has been hampered by the lack of a clear definition of the difference between teacher factors (effects) and teacher effectiveness, as well as confusion regarding the terms quality teaching and teaching effectiveness. Ding and Sherman's (2006) multilevel conceptual model provides both the needed definitions and the clarification of functions for these variables necessary for researching teacher effectiveness. The concept of teacher effectiveness, defined as being adept at raising test scores, has been studied for as long a there have been teachers, with varied results. Only recently have statistical methodologies have become more sophisticated and longitudinal data become more easily accessed, which allows researchers to better explore the variance in teacher effectiveness (Kane & Staiger, 2008b; McCaffrey, Sass, & Lockwood, 2009).

Just as research methods have evolved over the years, so has the teaching profession, which has become more concerned with quality and standardization. The process-product research of 1960s and 1970s and research reports in the 1980s lead to the development of the National Board for Professional Teaching (NBCT), whose mission is to increase teacher effectives throughout the United States by holding teachers to the same high standards as other professions. Notwithstanding NBCTs rise in popularity with policy makers, school administrators, and teachers unions, there was little empirical research to support their teacher effectiveness claims, which led to a myriad of studies being conducted.

Even with this flurry of research, only 30 empirical studies were found during the research and retrieval phase of the meta-analysis, and these provide conflicting evidence regarding NBPTS as a signal of effectiveness. The conflict may be because, although research has shown that teachers play the significant role in student achievement, students themselves and school or classroom contextual factors or both also influence student achievement. In addition, variances that are not attributable to teacher, student, and contextual effects may be the result of methodological variance. These factors are coded as possible moderator variables in the meta-analysis.

CHAPTER III

METHODOLOGY

The purpose of this study was to conduct a meta-analysis in order to aggregate the research findings of empirical studies that investigated the relationship between National Board Certification (NBC) status and student achievement and by analyzing moderator variables. This chapter contains the methodology of the study: the research design, literature search, including inclusion and exclusion criteria, coding procedures, study identification information, data analysis, and validity and reliability.

Research Design

According to Durlak (1995), the purpose of meta-analysis is to review quantitatively the results of a research domain with the intent of identifying any significant relationships that exist between study features and outcomes. Meta-analysis particularly is useful for drawing conclusions with more confidence when individual primary studies present conflicting findings (Cooper, 1998), as is the case with research regarding the National Board for Professional Teaching Standards (NBPTS).

In addition to comparing study features when conducting the data analysis, effect-size measures (the dependent variable) are used to compare two groups, which, in the current study, are National Board Certified Teachers (NBCTs) and teachers who are not National Board Certified (nonNBCTs). Calculating the average effect size across studies can result in increased sample size, statistical power, and consequently, the reliability of findings on the effect of National Board Certified Teachers (NBCTs) on academic achievement (Lipsey & Wilson, 2001). From its inception, the NBPTS has envisioned having a significant effect on teacher effectiveness and student learning. To evaluate this objective, numerous researchers have conducted empirical studies of the effect of NBC on student achievement using test scores as the outcome variable because they are the best quantitative measures available for statistical analysis (Hakel, Koenig, & Elliott, 2008). The findings of these studies have been inconsistent. Therefore, comparing the data reported in the various empirical studies using effect sizes as the common scale provides a statistical standardization of study results. This standardization provides numerical values that are interpretable in a consistent manner across all the variables and measures involved (Lipsey & Wilson, 2001).

A search of the literature on NBC and student achievement uncovered inconsistent quantitative results, and yet, the source of variance in these studies is unclear. Although variance can be the result of chance fluctuations in sampled estimates, the question of why obtained results vary across studies needs investigation. Using standard deviation as a measure of variability along with the corresponding mean gives only an indication of the variability of effect sizes (Durlak, 1995). For this reason, methodological, source, and study features were coded and analyzed in order to account for all meaningful study differences. Only achievement outcomes were analyzed because a review of the literature by the Hakel et al. (2008), found that all of the empirical studies that have been done to evaluate the effectiveness of NBC compare the achievement scores of students taught by NBCTs and nonboard certified teachers.

Literature Search

The general approach to the identification and selection of empirical literature relevant to the effect of National Board Certification on student outcomes for this synthesis was to start with broad categories and many search terms and then progressively narrow the group of studies down to those appropriate for the application of meta-analytic techniques.

Data Sources

First, published works were retrieved through the use of comprehensive sources traditionally used for conducting a literature search. These sources included PsycINFO, the Educational Resource Information Center (ERIC), ProQuest, Dissertations: Dissertation Abstracts International (DAI) and Dissertation Abstracts Online (DAO), Sage On-Line Publications, and Google Scholar. Next, the NBPTS Research Board website was explored through a link to relevant studies. Studies that could not be obtained from these sources were retrieved using Sage Publications Online, Google Scholar, Link+, and interlibrary loans.

Search Strategies

To find studies not produced through these search methods, the reference lists of the studies that were found were used to search for others. Manual searches of educational journals and professional journals from a personal library also were conducted. In an attempt to locate unpublished works, personal letters were sent to experts, Misty Sato, Lee and Judy Shulman, and Linda Darling-Hammond, who might have had knowledge of the existence of additional studies. Personal contacts, including National Board Certification friends from the University of San Francisco and Stanford University, also were contacted for recommendations. Although this list is extensive, it is not exhaustive.

Keywords that were used to search databases and other related queries included National Board Certification (NBC), NBC and quantitative, NBC and empirical, NBC and student outcomes, NBC and achievement, NBC and teacher quality, NBC and teacher expertise, NBC and teacher effectiveness, NBC and professional development, National Board for Professional Teaching Standards (NBPTS), NBPTS and quantitative, NBPTS and empirical, NBPTS and student outcomes, NBPTS and achievement, NBPTS and teacher quality, NBPTS and teacher expertise, NBPTS and teacher effectiveness, and NBPTS and professional development. The use of these keywords resulted in a master candidate list of 30 studies that appeared to meet the inclusion criteria based on the title and abstract. Further examination of the methods and results sections narrowed the master candidate list to 22 studies. These 22 studies were then coded based on the following inclusion and exclusion criteria, which reduced the master list to the final 12 studies included in the current meta-analysis.

Inclusion and Exclusion Criteria

Studies included in this investigation and those that were excluded were identified from the master candidate list based on written specifications. The written specifications for the criteria a research study had to meet in order for the findings to be included in this meta-analysis are listed below.

1. Studies had to be conducted in the United States between 1994 (the first year for the NBPTS assessment process) and 2009.
2. Outcomes needed to be represented in terms of standardized test scores for reading and mathematics. One study that focused on high-school science and one study that pertained to high school vocational training were both included with the mathematical studies because of the mathematics content covered in those subjects.

3. Only studies examining 3rd through 12th grade teachers were included. Although there is an Early Childhood Generalist and an Exceptional Needs certification that covers prekindergarten through second grade, current empirical research using standardized testing begins in the third grade when statewide testing begins.
4. Studies had to examine whether there were differences in test scores for students of board certified teachers and students of nonboard certified peers.
5. Studies needed to provide sufficient data to compute an effect size between NBCTs and nonNBCTs.

The written specifications for the criteria used to determine the exclusion of findings from a research study in this meta-analysis are listed below.

1. Qualitative research and case studies were not included because they did not include descriptive statistics.
2. Studies that used value-added designs that did not provide sufficient data to compute an effect size between NBCTs and nonNBCTs were excluded.
3. Empirical studies that examined whether board certified teachers were more or less effective than their nonboard certified peers or themselves by comparing standardized test scores for the year before they certified, the year they certified, and the year after they certified were excluded.

Prior to the coding process, eight studies were excluded from the master list of 30 based on the fact that they examined whether board certified teachers were more or less effective than their nonboard certified peers or themselves by comparing standardized test scores for the years before, during, and after they certified. During the coding process, 10 studies were excluded. Table 2 contains the primary author's name, the year of publication, and the publication type for research excluded from this meta-analysis during the coding process. The reasons for exclusion are also presented in the table.

Table 2
Excluded Studies by Primary Researcher with Reasons for Exclusion from the Master Candidate List of 22 Studies

		Reasons	
Primary Researcher	Year	Missing Data	Incompatible Statistics
Angle	2006	X	
Bundy	2006	X	
Cantrell	2007	X	
Clotfelter	2006	X	
Harris	2007		X
Holland	2006		X
Kane	2008	X	
Sanders	2005	X	
Stone	2002		X
Vitale	2008	X	

Coding

This section pertains to the researcher-developed coding instrument, pilot testing, and information regarding study coders, including their qualifications and training. Procedures for ensuring interrater reliability also are described.

Coding Protocol

After studies had been identified and collected, the meta-analysis information was extracted using the coding protocol (Lipsey & Wilson, 2001). Accordingly, a coding protocol and coding manual were developed so that studies would be coded based on study descriptors, sample, methodology, and moderator variables. The moderator variable is a characteristic from a reviewed study that accounts for significant variability in effect sizes there should. Therefore, when a possible moderator variable differs between studies, there should be a difference in the magnitude of their effect sizes (Durlak, 1995). In order to decide which study characteristics to include as moderator variables, studies were evaluated for factors found and discussed in chapter II regarding teacher, classroom, and school factors on student learning gains (Burke & Sass, 2008; Jargowsky & El Komi, 2009; Odden, Borman, &. Fermanich, 2004; Sanders & Rivers, 1996; Veldman & Brophy, 1974).

The coding process began with the researcher reviewing a sample set of studies to learn of the types of variables common to most studies selected for the meta-analysis. Coders then reviewed the studies that were chosen and completed a coding protocol. At this initial stage, when the instrument was developed, information that had potential for use in later analysis was included for coding. Consequently, not all coded information was used in the final analysis. The coding protocol is located in Appendix A.

Coders

Because the researcher is a National Board Certified teacher, there might be a potential for bias in coding (Cooper, 1998). Thus, objectivity of the coding process was ensured by the use of multiple coders, two of whom had no experience with or connection to the NBC process. The two additional coders, who are graduate students selected due to familiarity with effect sizes and other descriptive statistics, were trained to code the 22 original studies.

To ensure the reliability of coding procedures prior to pretesting the coding protocol, the researcher trained two coders so that they would be familiar with the coding sheet and the coding manual (see Appendix A; Cooper, 1998). During the first meeting, the researcher explained the purpose and rationale of the study and reviewed the coding manual that contained an explanation of each feature of the coding sheet. Training included careful reading and discussion of the coding protocol, manual, and response options. Use of the coding sheet for collecting data from a primary study also was demonstrated, and any issues that were not clear to the coders were discussed.

Pilot Testing and Interrater Reliability

A pilot test of the coding protocol was conducted in order to assess the usefulness of the coding sheet and to assess whether it required refinement (Cooper, 1998). A randomly selected study was chosen for this purpose, was coded by the researcher and one other trained coder, and results were compared. Information acquired during subsequent discussions regarding the functionality of the coding protocol was used to revise category descriptions before a second study was coded in order to ensure the reliability of the instrument. After revisions were made to the coding protocol another trained coder and the original two coded a second study. After comparing results and discussing need refinements, a second revision of the coding sheet was completed.

Once the evidence of reliability of the coding sheet was confirmed through pilot testing, interrater reliability was checked by comparing the coding results of a third study coded by the researcher and each of the other coders. A percentage agreement procedure was used to assess coding reliability (Lipsey & Wilson, 2001). The agreement rate was calculated by dividing the total number of agreed-on codings by the total number of codings (Cooper, 1998). Any disagreements in coding were discussed and resolved by consensus. There was 92% interrater reliability. Coding of the fourth study by all three coders resulted in 96% interrater reliability, and coding differences were resolved by consensus. At this point, coders began coding randomly assigned studies. Two of the remaining 18 studies were coded by one coder before that coder was no longer available. The other 16 were divided between the researcher and the remaining coder.

Coding Categories

Study descriptors that account for different results across studies were coded in three areas: (a) study identification consisting of the publication type and year of publication, (b) outcome measures comprised of general study context information that was valuable for descriptive purposes and statistical calculations, and (c) coded variables: dependent, independent, and moderator variables studied in this meta-analysis.

Study Identification Information

Coded publication data included an American Psychological Association (APA; 2009) citation for the study, the publication or reporting year, the type of publication, and publication title by initials only. Also included was the database or source where the study was obtained.

Coded Variables

The dependent variable coded for the study was student achievement in reading and mathematics or vocational training, indicated by differences based on end-of-year or end-of-instruction test scores. Assessment measure descriptors included pre- and posttest gain scores and end-of-course grades.

The coded independent variable had two levels based on the teachers' certification status. The first level was National Board Certified Teacher (NBCT) and the second was nonNBCT (teachers who were not board certified).

Also coded were special independent variables and additional study characteristics that were hypothesized to affect findings. These additional characteristics are referred to as moderator variables and are discussed in more detail in the next section of this chapter (Rosenthal, 1991). Using the Ding and Sherman's (2006) conceptual framework for effective teaching discussed in chapter I and a review of the literature on factors associated with variations in the magnitudes of relationships between effective teaching and student achievement (Burke & Sass, 2008; Jargowsky & El Komi, 2009; Odden et al., 2004; Rosenthal, 1991; Sanders & Rivers, 1996; Veldman & Brophy, 1974), each study was inspected for variables that could moderate or change the relationship between certification status and student achievement.

The moderator variables coded for studies in this meta-analysis are presented in Table 3. The frequencies listed in the table were used to decide if sufficient data were available to explore whether effect sizes were moderated by these variables. Variables where the categories sum to more than 12 are the result of studies that included more than one category. The following sections contain details of each variable type.

Table 3
Variables Coded for Meta-analysis with Frequency of Study

Variable	Levels of the Variables	Frequencies
Publication Type	Dissertation	6
	Published	6
Subject	Mathematics	12
	Reading	9
School Level	Elementary	9
	Middle School	2
	High School	3
Student	Gender	12
	FRL*	6
	Ethnicity	3
	English Language Proficiency	3
Teacher	Ethnicity	7
	Years of Teaching	6
Assessments	Pretests	8
	No Pretest	4
Assessment Type	Criterion Referenced	4
	Norm Referenced	8

*FRL = Free-and-Reduced Lunch

Teacher Characteristics

Based on a review of the literature, the possible teacher characteristics that might mediate effect sizes include licensure, number of years teaching, additional degrees, type of program being taught (special education, gifted and talented education (GATE), or general education), gender, and ethnicity and race. The possible moderator variables coded for teachers that were considered for analysis in this study are presented in Table 3. Those variables that are not included in Table 3 were not included by the primary researchers in their studies.

Student Characteristics

The possible moderator variables that are included in the literature as having an effect on research outcomes included student ages, race, gender, socioeconomic status based on participation in free-and-reduce lunch (FRL) programs, grade level, English Language Proficiency, placement in programs for students with exceptional needs, and parent education levels. The possible moderator variables coded for students that were considered for analysis in this study are shown in Table 3. As with teacher characteristics, those variables that are not included in Table 3 were not included by the primary researchers in their studies.

School Contextual Characteristics

According to the literature, the influence of economic status and other background variables are considered to be possible moderator variables that should be controlled. Unfortunately, the review of the sample studies and the final analysis of the research did not provide adequate numbers of contextual characteristics to warrant coding for them, and they are not included in the table.

Methodological Quality

Two areas of coding reliability were discussed in earlier subsections. The third area of coding addressed to ensure internal validity was establishing the importance of substantive and methodological features. The second chapter of this study provided a review of the literature on teacher, student, and school effects; teacher effectiveness; and value-added methods; process-product approaches; standardized assessment; and other factors that influence student outcomes on the achievement measures. The chapter also covered methodological features that have been shown to influence the results of empirical studies including type of assignment, methods of data collection and analysis, and the type of assessments that provided the achievement results.

To further establish the importance of methodological features, methodological quality as it relates to effect-size was coded as a moderator variable using degree of rigor (Durlak, 1995). Rigor was defined as using certain types of sample assignment, using more rigorous data analysis methods, and the type of assessment data used to calculate the effect size. Data-analysis types used for quality evaluation included hierarchical linear modeling (HLM), ordinary least squares (OLS), independent-samples *t* test, analysis of variance (ANOVA), analysis of covariance (ANCOVA), and multivariate analysis of variance (MANOVA). The assignment types included in the quality analysis were matching, blocking, and random or nonrandom assignment. A final area coded for methodological quality was the use of a pretest for equivalency. The study quality variables that were considered as possible moderators are listed in Table 4.

Methodological characteristics of each study were varied. For instance, only 31% of the studies used hierarchical and regression methods to analyze student outcomes with

Table 4
Coded Variables for Study Quality with Frequency of Study

Variable	Levels of the Variables	Frequency
Methodology	Regression	2
	t test	3
	ANOVA	3
	MANOVA	1
	HLM/OLM	3
	ANCOVA	1
Assignment	Matching/ Blocking	7
	Nonrandom	6
Design	Pretest	6
	No Pretest	7

the remaining studies using *t* tests or some form of analysis of variance. Methodological study features are presented in Table 4. The other two areas of methodological quality examined as possible moderator variables in this meta-analysis were including a pretest in the study and adjusting for the nonrandom placement of students and teachers in classrooms by using matching or blocking of the groups being compared were evenly divided between studies.

Description of Studies Included in the Meta-Analysis

Based upon the inclusion and exclusion criteria, a total of 12 studies could be used for data analysis after all 22 studies had been coded. The primary author’s name, year of publication, and the publication type for research included in this meta-analysis are shown in Table 5. All included reports were published between 2004 and 2007.

There were an almost equal number of published and unpublished studies included in the meta-analysis, as shown in Tables 5 and 6 on the following page, with the majority of the studies being completed between 2004 and 2007. All of the studies included mathematics, but only 69% included reading. The majority of the studies were

Table 5
Reading Studies by Primary Researcher, Year and Type of Publication, Assessment, Study Size, and Assessment Type

Primary Researcher	Year	Publication Type	School Level*	Assessments	Assessment Type
Benigno	2005	Dissertation	ES, MS,	Pretest	Criterion Referenced
Childs	2006	Dissertation	ES,	No Pretest	Criterion Referenced
Falaney	2006	Dissertation	ES,	Pretest	Norm Referenced
Fisher	2005	Published	ES, MS	No Pretest	Criterion Referenced
Goldhaber	2004	Published	ES,	Pretest	Norm Referenced
McCloskey	2005	Published	ES,	Pretest	Norm Referenced
Rouse	2004	Dissertation	**	No Pretest	Norm Referenced
Silver	2007	Dissertation	ES,	Pretest	Norm Referenced
Vandervoort	2004	Published	ES, MS,	Pretest	Norm Referenced

*Elementary School (ES), Middle School (MS) and High School (HS)
** Effect sizes for Rouse combined for 3rd-8th grade and could not be disaggregated

Table 6
Mathematics Studies by Primary Researcher, Year and Type of Publication, Assessment, Study Size, and Assessment Type

Primary Researcher	Year	Publication Type	School Level	Assessments	Assessment Type
Benigno	2005	Dissertation	ES,	Pretest	Criterion Referenced
Cavaluzza	2004	Published	HS	Pretest	Norm Referenced
Childs	2006	Dissertation	ES,	No Pretest	Criterion Referenced
Falaney	2006	Dissertation	ES,	Pretest	Norm Referenced
Fisher	2005	Published	ES, MS	No Pretest	Criterion Referenced
Goldhaber	2004	Published	ES,	Pretest	Norm Referenced
McCloskey	2005	Published	ES,	Pretest	Norm Referenced
Rouse	2004	Dissertation	HS	No Pretest	Norm Referenced
Rouse/Hollomon	2005	Published	HS	No Pretest	Criterion Referenced
Silver	2007	Dissertation	ES,	Pretest	Criterion Referenced
Stephens	2003	Dissertation	ES,	Pretest	Norm Referenced
Vandervoort	2004	Published	ES, MS	Pretest	Norm Referenced

conducted using data from elementary-school test scores, with only 15% using middle-school test scores in mathematics, and 23% using reading test scores. For high school, there were no studies for reading, and 31% of the studies used data from high-school mathematics assessments. Only 30% of the included studies used criterion-referenced tests, and only 30% of the studies utilized a pretest. Finally, very few of the studies provided demographic data, which precluded analysis of variables that may have moderated the outcomes of the studies (Table 3, p. 65).

Two comparison groups were defined as Nationally Board Certified Teachers (NBCTs) and nonBoard Certified Teachers. All student sample participants ranged in age from 8 to 18 years old and were enrolled in 3rd through 12th grade. The majority of the children (38%) took their standardized test in North Carolina, which is one of the states with a large cadre of Nationally Board Certified teachers. South Carolina is the state with the second largest number of studies, reflecting 23% of the research, and Florida represented another 15% of the studies. The remaining studies were evenly divided between Mississippi, Oklahoma, and Arizona. The Arizona study included results from the entire state, whereas the majority of the remaining studies used data from schools in urban and metropolitan areas.

Data Analysis

The following section contains the data-analysis procedures used to analyze information collected from the primary studies. First is an explanation of the effect-size measure, and then specific procedures are described for each research question.

Effect-size Measure

Two-variable relationships constitute the most typical type of research findings that are most commonly meta-analyzed (Lipsey & Wilson, 2001). In the present study, the particular form of the relationship is a group contrast in which student outcomes of NBCTs and nonNBCTs are compared. The dependent variable used for contrasting the two groups is standardized test scores analyzed separately for reading and mathematics.

One tenant of meta-analysis is that there can be no more than one effect size per construct per study so the first step was to create an independent set of relevant effect sizes. Because most of the studies provided multiple means, standard deviations, and student sample sizes, it was necessary to pool both the means and the standard deviations, separately in order to calculate the single effect size for the study. Means for both NBCTs and nonNBCTs were aggregated separately for mathematics and reading by multiplying the number of students in each group by the individual group means and then dividing this sum by the total number of students in all groups (Hinkle, Wiersms, & Jurs, 2003). Next, standard deviations were pooled in order to calculate the standardized mean difference effect size statistic (Lipsey & Wilson, 2001, p. 48).

The effect-size statistic that is most appropriate for research findings in the form of group contrasts that are presented as the differences between mean values on student outcomes for NBCTs and nonNBCTs is the standardized mean difference. Thus, the next step was calculating the difference using the pooled means and standard deviations reported in primary research studies (Lipsey & Wilson, 2001, p. 48).

The d index has been shown to be upwardly biased based on small sample sizes. Hedges’s *g* is defined as a variation on Cohen's *d* that corrects for biases due to small sample sizes. Hedge’s formulae (Lipsey & Wilson, 2001, p. 49), therefore, were used for bias correction, and all subsequent computations used the corrected (unbiased) effect size.

When studies failed to report the means and standard deviations, every effort was made to contact the researcher(s) to retrieve the missing information directly. In the cases where it was not possible to retrieve missing information, the effect sizes were estimated. One study provided sufficient statistical data to use a different formula for estimating the effect size. The *F* ratio from a one-way analysis of variance and total sample size were used to calculate the standardized mean difference effect-size statistic separately for mathematics and reading in a study conducted by Childs (2006). The study had equal sample sizes as required for this calculation (Lipsey & Wilson, 2001, p. 199, formula 5).

The effect sizes from all studies included in this meta-analysis were averaged to compute an overall mean effect size of the relationship between certification status and student compute. The mean effect size was computed by weighting each effect size by the inverse of its variance (Lipsey & Wilson, 2001).

Test of Homogeneity

Hedge's *Q* (Hedges & Olkin, 1985), a test of homogeneity, was used to analyze the distribution of effect sizes of achievement outcomes for both mathematics and reading separately. The *Q* statistic was used to assess whether the variance in factors produced by the included studies was statistically significantly different from sampling error (Cooper, 1998). Hedge's *Q* statistic is distributed as a chi-square; therefore, if the effect sizes were not homogeneous, as indicated by a critical value that exceeded the critical value of a chi-square statistic with *k*-1 degrees of freedom, then homogeneity was rejected and extreme effect sizes were eliminated to obtain a homogeneous set of studies and the overall effect size. Assuming the calculated *Q* statistic for the individual effect sizes exceeded the critical value for the upper limits of the chi-square distribution, individual effect sizes were pooled and an average effect size was reported.

Data-analysis Procedures by Research Question

The meta-analysis specifically examined three research questions. Each question is discussed below in order to organize the explanation of the data-analysis procedures.

Research Question #1:

What is the effect on student achievement in mathematics and reading for students taught by teachers with National Board Certification (NBC) when compared with students taught by teachers without NBC?

The first research question is related to student outcomes based on a teacher's board certification status. After effect sizes were calculated for each study using the common index, Hedge's *g,* they were interpreted. The standards for the evaluation were a *g* of .20 is a small effect, a *g* .50 is a medium effect, and a *g* of .80 is a large effect, which have been found by Valentine and Cooper (2003) to be uncharacteristic of the magnitude of effect sizes found in educational research. Therefore, to assess the magnitude of research findings in this meta-analysis, Hedges and Hedberg's (2007) benchmark was used. For this benchmark, effect sizes near .20 are an important outcome when they are based on measures of academic achievement. Confidence Intervals (CIs) were generated around the average effect-size estimates for mathematics and reading separately.

Research Question #2:

What is the difference in the effect size of reading and mathematics assessments for *students taught by NBCTs when compared with students taught by nonNBCTs?*

The second research question also used the confidence intervals calculated for the first research question. The CIs for mathematics and reading were inspected for overlapping in order to assess differences between the averages measured for both subjects. Using confidence intervals is a fairly conservative way to test for differences and is appropriate because the effect size for both subjects were calculated using the same samples violating the assumption of independence.

Research Question #3:

To what extent do variables, such as type of assessment, type of publication, study size, and study quality moderate the relationship between Certification Status and academic achievement?

The final question explored whether there are differences between the groups in student outcomes for mathematics and reading related to moderator variables. Analog to analysis of variance (ANOVA) was used to test the difference between categories of moderator variables for the effect sizes (Cooper, 1998; Lipsey & Wilson, 2001). This method groups effect sizes into mutually exclusive categories. If the between category variance is statistically significant, the mean effect size across groups differ by more than the sampling error. This information was used in assessing the adequacy of the moderator variable in explaining the original heterogeneity among effect sizes (Lipsey & Wilson, 2001).

Reliability and Validity

The literature on validity and reliability in meta-analyses is discussed in terms of trustworthiness (Cooper, 1998; Cooper & Hedges 1994). Numerous steps were taken in the current meta-analysis to ensure the trustworthiness of the findings. These procedures for enhancing the validity and reliability of the current meta-analysis as well as for examining and reducing bias are the focus of this section.

External Validity

One threat to external validity from primary studies was the problem of study quality related to the reporting of study statistics. More than half of the excluded empirical studies were eliminated, at least impart, because of failure to report adequately descriptive statistics that would allow the computation of standardized mean differences. Plus, the studies that were included either did not use or did not report the results of rigorous methods that took into account variables presented in chapter II that research has shown to have a substantial effect on study results. Consequently, this lack of information may have influenced validity of this meta-analysis. In an attempt to minimize this threat to validity, researchers were contacted by phone or email with a request to supply any available missing data. Although several researchers responded to requests for information, none provided the requested data.

Publication Bias

There are two sources of publication bias that can affect the reliability of a meta-analysis. The first is the *file drawer effect* and the second one is *fail-safe N*.

File Drawer Effect

There is a tendency on the part of research publications to favor empirical studies reporting statistically significant effects and to deny publication to studies finding no statistically significant relationship between variables. The practice of reporting and publishing only positive outcome research creates a misrepresentation of the subject under investigation, especially if a meta-analysis is to be performed (Lipsey & Wilson, 2001). To address this problem – called the file drawer effect because researchers file away studies that find nothing of statistical significance or causal consequence (Rosenthal, 1991) – every effort was made to locate research findings that were not in journals (Light & Pillemer, 1984). These efforts included contacting prominent researchers in the area of NBC and student outcomes in order to request their unpublished work and referrals to others with unpublished manuscripts. Information regarding websites and listservs from organizations, such as the American Educational Research Association, the American Psychological Association, and WestEd, was sought in order to obtain papers or names and contact information of presenters**.** None of the correspondences received replies and a search of the listserv for WestEd did not yield any candidates for the master list.

Fail-safe N

The statistical significance of the results of a study is often the determinate in publication decisions. This means that smaller samples with lower statistical power may be underrepresented in the current study. Difficult-to-find studies have a strong effect on a meta-analysis because, as a synthesis of literature on a particular topic, their absence can cause an upward bias of the mean effect size. The lack of nonsignificant findings also influences statistical significance testing because it is based on the availability of only articles that produce significant findings. To address this problem Rosenthal (1991) developed the *fail-safe N* statistic to estimate the number of unpublished or unretrieved studies showing a zero effect that would need to be found in order to make the effects of the meta-analysis nonsignificant (Long, 2001). In addition to an exhaustive search of the literature to combat this threat, Rosenthal's method of computing *fail-safe* N was used to test if the overall effect size was statistically different from zero (Cooper, 1998; Lipsey & Wilson, 2001).

Assumptions

Independence is the statistical assumption that groups, samples, or other studies in the meta-analysis are unaffected by each other (Cooper, 1998). The assumption of independence for comparisons and relationship strength was met because students took the test independently and samples used to calculate a statistic were independent of each other. Nonindependence was still a minor issue because mathematics and reading scores, although analyzed separately, were taken from the same populations. Likewise, they shared historical and situational influences (Cooper, 1998). This nonindependence was an issue in analyzing data regarding the difference in the effect-size of reading and mathematics assessments for students taught by NBCTs when compared with students taught by nonNBCTs for research question 2. Accordingly, the conservative method of using CIs was the method used to explore differences related to subject.

Summary

This chapter contained an examination of techniques applicable to properly interpreting the results of this meta-analysis. First, how studies were located and what criteria were used to determine if the studies found in the search were included or excluded were discussed. Next, there was an explanation of the design of the coding scheme that was used to record the relevant data from each study to be included in this meta-analysis. This discussion was followed by a review of the dependent, independent, and moderator variables being studied, and details of effect size calculations were provided. Finally, there was an account of the techniques used for analyzing the data.

CHAPTER IV

RESULTS

The purpose of this study was to conduct a meta-analysis in order to examine the aggregated research findings of empirical studies that explored the relationship between National Board Certification (NBC) status and student achievement. In addition, moderator variables were analyzed to explore the possibility of differences in study features, explaining inconsistencies across studies. Results are presented in order for each research question.

Statistical Analysis

The results of the meta-analysis detailed in this chapter are derived from a literature search that culled 30 studies from four electronic databases. Out of those 30 reports, 12 met the criteria for inclusion in the meta-analysis. In total, the 12 studies yielded 21 independent effect sizes based on standardized mathematics and reading test scores. Descriptive statistics and the analog to analysis of variance (AVOVA) were used in order to investigate if statistically significant differences exist between the two comparison groups, which were defined as Nationally Board Certified Teachers (NBCTs) and nonBoard Certified Teachers (nonNBCTs).

Research Question 1

> *What is the effect on student achievement for students taught by teachers with National Board Certification (NBC) when compared with students taught by teachers who are not board certified?*

The first question investigated the effect of certification status on student achievement for both mathematics and reading. Overall, there was a combined total of

1,962,044 mathematics and reading achievement scores used for these analyses. Table 7 displays the means and standard deviations for reading achievement scores for both NBCTs and nonNBCTs. These variables were used to investigate whether having a teacher with National Board Certification (NBC) had a positive influence on students' reading achievement scores. Box-and-whisker plots were used to graphically display the contrast and degree of dispersion of the effect sizes from NBCTs and nonNBCTs for the 12 studies in the meta-analysis. An inspection of these displays gave a visual presentation of the effect sizes for both groups. The box plots in conjunction with the *Q* statistic were then used to contrast both the central tendency and dispersion of effect sizes (Lipsey & Wilson, 2001, p. 144).

Table 7

Sample Size, Mean, Standard Deviation, and Hedge's *g* for Student Achievement Scores in Reading Reported by Primary Researcher, Certification Status, and Grade Level

		Certification Status						
		NBCT	Non NBCT	NBCT	Non NBCT	NBCT	Non NBCT	
Primary Researcher	Variable	*n*	*n*	*M*	*M*	*SD*	*SD*	Hedge's *g***
Benigno	Grade 3 & 6	131	56	12.50	14.34	51.24	42.11	-.04
Childs*	Grade 4 & 5	530	525					.06
Falaney	Grade 4 & 5	525	567	14.03	13.88	25.08	28.48	.02
Fisher	Grade 4-8	25665	11329	590.22	590.19	13.40	13.93	.03
Goldhaber	Grades 3-5	600261	4297	6.18	5.69	6.37	6.13	.08
McClosky	Grade 5	4215	417	0.03	-0.01	1.05	0.99	.04
Rouse	Grade 3-8	369	415	3.44	3.39	0.74	0.72	.07
Silver	Grade 3-5	4572	4572	88.37	85.51	10.23	13.20	.24
Vandevoort	Grade 3-6	243874	1777	25.94	21.87	23.79	23.80	.16

*Childs' ES was calculated using the reported *F* value.

**Hedge's *g* was calculated from developmental scale scores, means, mean gain scores, gain scores, proficiency scores, adjusted mean gain scores, scale scores, and residuals from standardized achievement tests.

The means of the NBC teachers were higher than means for nonNBC teachers except for Benigno’s study, which found that students of nonNBCT had higher achievement scores than students of NBCTs. Additionally, the means for NBCTs in the Fisher, McClosky and Rouse studies are only slightly higher than the means for the nonNBCTs. An analysis of the means for both the NBCT and nonNBCT student outcomes indicate that, on average, students of NBCTs had higher reading achievement scores than students whose teachers were not board certified.

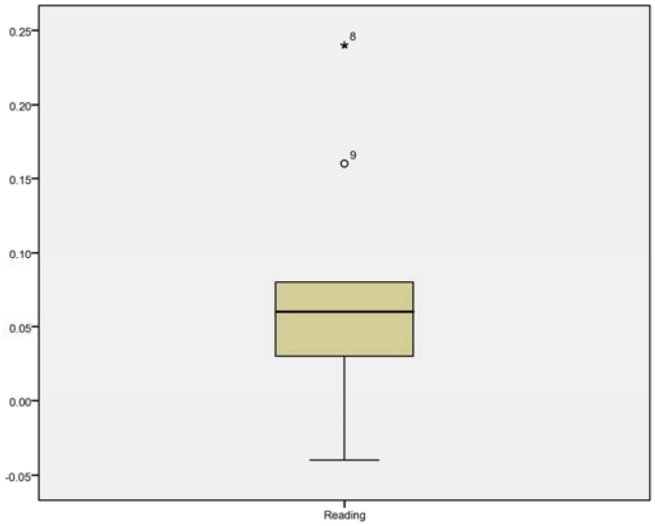


Figure 2. Box plot of Hedge’s *g* for Reading Achievement Scores

A commonly used effect size index for two group comparison studies is the standard mean difference. Therefore, effect sizes were calculated by finding the standardized mean difference while employing the Hedge’s *g* correction for sample bias and inverse variance weights (Lipsey & Wilson, 2001). All except one effect size were calculated from the reported means and standard deviations. The effect size for Childs’ study was calculated from the *F* ratio. Effect-size calculations were such that positive values indicated that student test scores for NBCTs showed greater improvement than nonNBCTs (Lipsey & Wilson).

There were nine effect sizes representing 942,370 reading achievement scores. The effect sizes presented in Table 7 and displayed in Figure 2 range from -.04 to .24 with approximately 89% being positive, indicating that NBCTs have a positive effect on the achievement scores of their students. This positive effect, however, is minor given that the overall weighted mean effect size of .09 is not statistically significant (see Table 8). The confidence interval presented in Table 8 ranges from .07 to .10 and does not contain zero, indicating there is relationship between student achievement of NBCTs and student achievement of nonNBCTs. Because the confidence interval does not contain zero, the mean effect size is statistically significant. An inspection of the standard deviations and the box plot in Figure 2 indicates that there is very little variance between the effect sizes with 75% of the values being between -.04 and .10. The remaining 25% of the values are spread out between .15 and .25.

Table 8

Overall Weighted Mean Effect Size, Homogeneity Statistic and Confidence Interval for Reading Achievement Outcomes

		95% CI			
Outcome	Mean ES	Lower	Upper	*Q* value	*df*
k=9, *N*=942,370	.09	.08	.10	.01	8

In addition to the overall weighted mean effect size for reading achievement, homogeneity of effect sizes was tested using Hedges and Olkin's (1985) homogeneity statistics *Q*. Like the box plot and the weighted mean effect size, the *Q* was not statistically significant, indicating homogeneity of effect sizes. Although there was homogeneity, two effect sizes in Table 7, Silver and Vandervoort, have large enough magnitudes to be considered practically important, according to Valentine and Cooper

(2003). Valentine and Cooper stated that using Cohen’s benchmark of .20 as a descriptor of the small magnitude of an effect size is not appropriate in an area like education where the effect sizes may be smaller than other areas in the social sciences. Therefore, Hedges and Hedberg’s (2007) standard was used. Because the effect sizes in this meta-analysis are based on academic achievement, an effect size of .20 is considered an important outcome. These two effect sizes notwithstanding, the range of effect sizes makes it difficult to make a firm declaration that teachers who achieved NBC have a greater effect on student achievement in reading.

The means and standard deviations for mathematics achievement scores for both NBCTs and nonNBCTs are presented in Table 9. These variables were used to

Table 9
Sample Size, Mean, Standard Deviation, and Hedge’s *g* for Student Achievement Scores in Mathematics Reported by Primary Researcher, Certification Status, and Grade Level

		Certification Status						
		NBCT	Non NBCT	NBCT	NBCT	Non NBCT	Non NBCT	
Primary Researcher	Grades	*n*	*n*	*M*	*M*	*SD*	*SD*	Hedge’s *g***
Benigno	3-6	362	382	543.79	309.76	41.85	43.89	.29
Cavalluzza	9 & 10	3049	98801	2016.84	1856.55	182.45	215.27	.00
Childs*	4& 5	525	530					.35
Falaney	4 & 5	1092	1092	17.82	17.72	25.11	29.79	.00
Fischer	4-7	15548	49768	577.11	586.70	14.79	14.47	.03
Goldhaber	3-5	4318	602577	10.21	9.75	7.00	6.92	.07
McCloskey	5	417	4215	0.08	0.01	0.97	1.01	.07
Rouse	3-8 & 9-12	985	940	3.08	2.86	0.82	0.81	.28
Rouse/Hollomon	9-12	726	744	3.52	3.49	0.66	0.68	.04
Silver	3-5	4572	4572	93.37	91.88	7.69	9.33	.17
Stephens	4 &5	154	158	470.65	464.42	13.57	13.31	.46
Vandevoort	3-6	1719	250145	27.65	26.15	25.48	25.67	.12

*Childs’ ES was calculated using the reported *F* value.
**Hedge’s *g* was calculated from developmental scale scores, means, mean gain scores, gain scores, proficiency scores, adjusted mean gain scores, scale scores, and residuals from standardized achievement tests.

investigate whether having a teacher with NBC had a positive effect on student's mathematics achievement scores. The largest study used the records of 606,895 students and the smallest study included 312 student test scores.

As with the *Q* statistic for reading, an inspection of Table 10 shows that the *Q* statistic for mathematics achievement scores was not statistically significant, indicating that the effect sizes of the study samples included in the meta-analysis are homogeneous. Also, like the reading achievement scores, the mathematics mean of students taught by NBCT were higher than means for students taught by nonNBCTs, with the notable exception of Fisher's study. An analysis of the means for both the NBCT and nonNBCT student outcomes indicate that, on average, students of NBCTs had higher mathematics achievement scores than students whose teachers were not board certified.

Table 10

Overall Weighted Mean Effect Size, Homogeneity Statistic, and Confidence Interval for Mathematics Achievement Outcomes

		95% CI			
Outcome	Mean ES	Lower	Upper	*Q* value	*df*
k=12, *N*=1,047,391	.08	.07	.10	.007	11

There were 12 effect sizes representing 1,047,391 mathematics achievement scores. An overall weighted mean effect size of .08 was computed and is statistically significant because the confidence interval does not contain zero, indicating there is a difference in the relationship between certification type and student achievement. An inspection of the Hedges's *g* column of Table 9 supports the positive impact of NBC with all of the effect sizes being positive. The effect sizes presented in Table 9 and displayed in Figure 3 range from 0 to .46 with 50% of the values being close to zero indicating that there is only a small difference between the average achievement of students whose teachers are NBC and those whose teachers are not NBC. This lack of a statistically significant difference is supported by the graphic display of mathematics achievement scores in Figure 3, which indicates that there is both statistical significance and practical significance for only 33% of the effect sizes. The remaining 66% are not statistically and practically significant, indicating little to no relationship between certification status and student achievement.

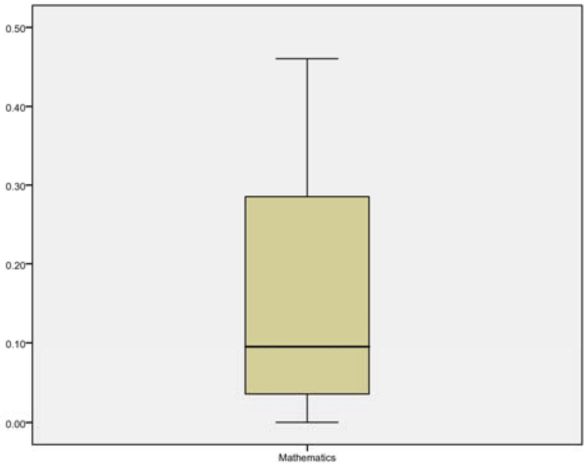


Figure 3. Box plot of Hedges's *g* for Mathematics Achievement Scores

Four of the effect sizes in Table 9 indicate that NBCTs have a greater effect on student mathematics achievement, and using Hedges and Hedberg's (2007) definition of magnitude, they are all statistically and practically significant. There are also two medium effect sizes showing that NBCTs have a more positive effect on students' mathematics achievement than nonNBCTs. In total, the effect sizes provide evidence that, on average, students of NBCTs outperform students whose teachers are not Board Certified, even though the mean effect size is small.

Research Question 2

What is the difference in the effect size of reading and mathematics assessments for students taught by NBCTs when compared with students taught by nonNBCTs?

The confidence intervals calculated for question one and the box plots were used to investigate whether there is a difference in the effect size of reading and mathematics assessments for students taught by NBCTs when compared with students taught by nonNBCTs. The confidence interval for the overall weighted mean effect size for reading was .08 to .10 and the confidence interval for the overall weighted mean effect size mathematics was .07 to .10. Because the confidence intervals overlap, there probably is no difference regarding the effect of NBC on student achievement (Cornell Statistical Consulting Unit, 2008). This lack of a difference between the effect of NBCTs for different mathematics and reading can be seen on inspection of the box-and-whisker plots in Figure 4. The majority of cases are close to zero for both subjects.

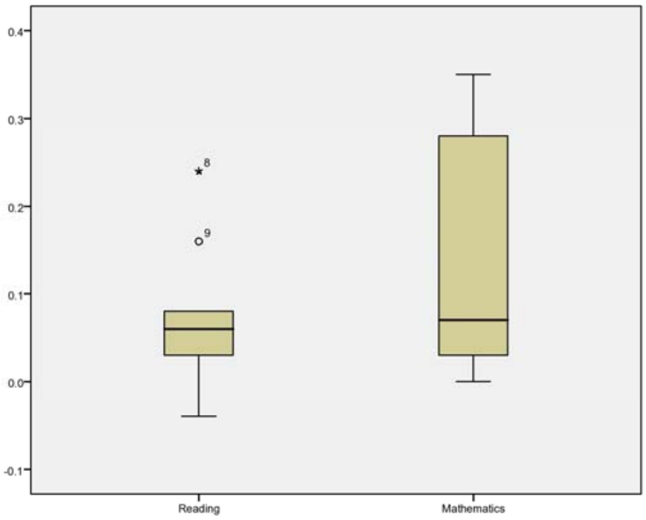


Figure 4. Comparison of Box plot of Hedges’s g for Reading and Mathematics Achievement Scores

Fail-Safe N

Rosenthal’s (1991) *fail-safe N* statistic was calculated to estimate the number of unpublished or not retrieved studies showing a zero effect that would need to be found in order to make the effects of the meta-analysis not statistically significant (Lipsey & Wilson, 2001). Because the number of published studies identified and available for the present meta-analysis was quite small, publication bias was potentially a threat to the external validity or generalizability of the findings with regard to the relationship between NBC and student achievement. Therefore, the *fail-safe N* statistic was conducted to assess the reliability of the meta-analysis. Using the effect size estimates from the 12 studies of mathematics, because mathematics had more studies than reading, the *fail-safe N* analysis was preformed employing the procedures developed by Rosenthal (1991). The results of the analysis indicate that approximately 9 null-result studies would be required to reduce the combined effect size to a statistically nonsignificant level. Given that a literature search of four major data bases produced only 30 studies that assessed the effect of NBC on student achievement scores, it seems unlikely that this many studies exist in researchers' file drawers. What may have created publication bias was the inability to incorporate studies that did not meet the criteria for inclusion due to missing data.

Research Question 3

> *To what extent do variables, such as type of assessment, kind of publication, study size, and study quality moderate the relationship between Certification Status and academic achievement?*

The analytic approach used to investigate the relationship between single categorical variables and the effect sizes of the studies included in this research study is an analysis of mean effect sizes by categorical study feature, analogous to a one-way analysis of variance (ANOVA). The findings from these studies presented in Table 7 and

Table 9 range from -.04 to .46. For this reason, moderator variables related to study characteristics (e.g., publication type, assessments, assessment type and school level) were analyzed using the analog to ANOVA to explore their influence on study outcomes. The *Q*-between (Q_B) value and *Q*-within (Q_W) value were calculated to estimate the homogeneity of results by mean effect sizes between and within variable and level categories, respectively. As presented in the definitions in chapter I in the analog ANOVA, the *Q* statistic from the analysis of variance is subdivided into $Q_{between}$, representing the variance in effect sizes accounted for by moderator variable, and a Q_{within}, representing within group error. When the $Q_{between}$ is statistically significant and the Q_{within} is not statistically significant, the moderator variable successfully accounts for the variability in effect sizes (Lipsey & Wilson, 2001).

The results by mean effect sizes of studies that explored the relationship between certification status and reading achievement are presented in Table 11. The results for the studies that explored the same relationship for mathematics achievement are presented in Table 12. An inspection of both tables indicates that all confidence intervals for each of the categories are narrow providing greater precision of the estimate of the mean effect size. Furthermore, except for middle school, all confidence intervals do not contain zero, signifying that the mean effect sizes are nonzero and statistically significant (Durlak, 1995; Lipsey & Wilson, 2001).

An examination of Tables 11 shows that all of the Q_B for all categories except Group Assignment are statistically significant, indicating the observed differences in both the relationship between certification status and reading may be statistically significant. This is not the case however. Although the individual Q_W s for school level, criterion referenced, Nonrandom, HLM/OLS/Regression, and No Pretest for reading are not statistically significant, after summing Q_W for to get the variance within each category, all of the Q_W were statistically significant. Given that the Q_B for Publication Type, School level, and Study Methodology were statistically significant and Assessment Type, Assessment, and Group Assignment are not statistically significant for reading none of the moderator variables for reading account for excess variability in the effect-size distribution.

Table 11

Results of Analog to ANOVA for Comparison of Moderator Variables in Studies Comparing NBCTs and nonNBCTs Reading Scores on Standardized Achievement Tests

				95% CI		
Variable and levels	Q_B	*k*	*g*+	Lower	Upper	Q_W
Publication Type	42.00*					
Dissertation		5	.19	.16	.23	22.84
Published		4	.06	.05	.08	24.47
School Level[a]	44.44*					
Elementary		8	.13	.11	.15	62.47
Middle School		3	.02	.00	.05	14.06
Assessment Type	40.80					
Norm Reference		6	.13	.11	.15	47.40
Criterion Reference		3	.03	.01	.06	.44
Study Methodology	58.77*					
t test/ANOVA/ ANCOVA/MANOVA		3	.18	.15	.21	24.00
HLM/OLS/Regression		6	.16	.14	.19	5.87
Group Assignment	1.65					
Matching/ Blocking		5	.08	.06	.10	76.89
Nonrandom		4	.10	.07	.12	8.80
Assessments	40.42					
Pretests		6	.13	.11	.15	47.79
No Pretest		3	.04	.01	.06	0.43

[a] Some studies provided data for more than one school level yielding more effect sizes (*k*=number of effect sizes) than studies.

[*] Statistically significant when the overall error rate was controlled at the .05 level.

Table 12

Results of Analog to ANOVA for Comparison of Moderator Variables in Studies Comparing NBCTs and nonNBCTs Mathematics Scores on Standardized Achievement Tests

				95% CI		
Variable and levels	Q_B	*k*	*g*+	Lower	Upper	Q_W
Publication Type	64.47*					
Dissertation		6	.20	.16	.23	26.79
Published		6	.05	.03	.06	17.16
School Level[a]	38.74*					
Elementary		9	.11	.09	.13	52.26
Middle School		2	.00	-.03	.03	3.96
High School		3	.04	.01	.07	27.33
Assessment Type	4.84*					
Norm Reference		7	.09	.07	.10	59.63
Criterion Reference		5	.05	.03	.08	43.95
Study Methodology	61.77*					
t test/ANOVA/ ANCOVA/MANOVA		8	.04	.02	.05	39.44
HLM/OLS/Regression		4	.16	.14	.19	7.20
Group Assignment	.89					
Matching/ Blocking		8	.07	.05	.09	103.92
Nonrandom		4	.08	.06	.11	7.29
Assessments	2.15*					
Pretests		8	.08	.06	.10	86.79
No Pretest		4	.06	.04	.08	49.60

[a] Some studies provided data for more than one school level yielding more effect sizes (*k*=number of effect sizes) than studies.

[*] Statistically significant when the overall error rate was controlled at the .05 level.

An inspection of Tables 12 shows that only the Q_B for all publication type, school level, and study methodology are statistically significant indicating the observed differences in both the relationship between certification status and mathematics may be statistically significant. Taken individually the separate Q_W s published, middle school, HLM/OLS/Regression, and Nonrandom are not statistically significant. As with reading, after summing Q_W to get the variance within a category, however all of the Q_W were statistically significant, and given that the Q_B for mathematics were also statistically significant publication type, school level, and study methodology and nonsignificant for group assignment, assessment and assessment type, none of the moderator variables for mathematics account for excess variability in the effect-size distribution.

Summary

The results of the literature search identified 12 studies that met the inclusion criteria for this meta-analysis that examined the effect of certification type on student achievement. Of those, 9 studies were used to calculate effect sizes for reading and all 12 were used to calculate the effect sizes for mathematics. The effect sizes obtained using Hedges's *g* ranged from -.04 to -.24 in reading and from 0 to .46 in mathematics.

The results of this meta-analysis regarding the first question investigating the differences in student achievement based on certification status were inconclusive. Although the results of this meta-analysis indicate that NBC has an influence on student achievement in both reading and mathematics, the observed differences in most cases are small and may not be statistically significant (Cornell Statistical Consulting Unit, 2008). Finally, the results for the third question examining moderator variables for possible contributions to variability show that none of the moderator variables for either reading or mathematics account for excess variability in the effect-size distribution. These results are discussed in chapter V.

CHAPTER V

DISCUSSION, LIMITATIONS, AND RECOMMENDATIONS

In the introductory chapter, the argument was presented that there is conflicting evidence regarding the effectiveness of National Board Certified teachers (NBCT) when compared with nonboard certified teachers, especially as measured by student achievement. To generate new evidence with regard to the relationship between certification status and student achievement, this meta-analysis analyzed moderator variables and examined the aggregated research findings of studies that explored the relationship between certification status and student achievement. This chapter includes a summary of the meta-analysis, an explanation of limitations likely to have influenced the results, and a discussion of the research questions with an interpretation of the results presented in the previous chapter. The chapter concludes with recommendations for research.

Summary of the Meta-analysis

An extensive review of the literature located 30 titles and abstracts related to empirical research on the difference between NBCTs and teachers who are not board certified (nonNBCTs). Studies that met the criteria had publication dates between 2003 and 2007 reflecting the fact that research into the effectiveness of National Board Certification is a recent field of study. Only 12 of the original 30 articles and nonpublished material met the inclusion criteria of (a) having outcomes that were represented in terms of standardized test scores for reading, mathematics, or mathematics-related courses, (b) being studies that quantitatively examined 3rd through 12th grade, (c) examining whether board certified teachers are more effective or less effective than their nonboard certified peers, and (d) provided sufficient data to compute an index of differences between NBCTs and nonNBCTs.

For collecting data from primary studies, a coding manual and coding form were designed corresponding to the multilevel education model of factors influencing student achievement presented in chapter I that is based on Ding and Sherman's (2006) multilevel dynamic education model. Then the descriptive data of six published studies and six dissertations on National Board Certification (NBC) and student achievement were examined using descriptive statistics to assess the comparative teaching outcomes of NBCTs and nonNBCTs for the purpose of creating generalizations. In addition, outcomes for these two categories of teachers were assessed across subject matter using descriptive statistics. Finally, study characteristics associated with differences in effect sizes were explored using analog to analysis of variance (ANOVA) in order to search for influences on previous findings in order to resolve conflicts in the literature.

As with the primary studies used for this meta-analysis, the findings indicate that NBC has an influence on student achievement in both reading and mathematics, however, the observed differences in most cases are small and may not be statistically significant (Cornell Statistical Consulting Unit, 2008). The results of this meta-analysis regarding the first question, investigating the differences in student achievement based on certification status, were mean effect sizes of .09 for reading and .08 for mathematics. This small difference in mean effect sizes for mathematics and reading also is reflected in the results for the second research question, a comparison of the confidence interval to explore the differences in achievement scores for both subjects between the students taught by NBCTs and with students taught by nonNBCTs. Finally, the results for the third question, examining moderator variables for possible contributions to variability, show that none of the moderator variables investigated for either reading or mathematics account for excess variability in the effect-size distribution.

Limitations

There are limitations unique to this meta-analysis given the criteria for inclusion and coded characteristics of the studies utilized. Three important limitations are given in the following subsections: (a) the limitations of using standardized achievement scores as a measure of teaching effectiveness, (b) missing data regarding possible moderator variables, and (c) the constraints of value-added methods.

Using Standardized Achievement Scores as a Measure of Teaching Effectiveness

The results of this meta-analysis are based on standardized achievement scores, which may provide a misleading estimate of National Board Certified Teacher effectiveness because of validity issues. Validity refers to how appropriate and meaningful the inferences are that can be drawn from assessment results based on their intended use. Two areas of concern of validity related to this body of research are using tests as a signal of teacher effectiveness and test quality. Michael Kane (2005) defined high-stakes testing by the criteria that the assessments are used to make decisions that come with consequences. Under No Child Left Behind legislation, the use of assessment is front and center in evaluating teacher effectiveness; however, there are many threats to validity in high-stakes achievement testing (Haladyna & Downing, 2004) that indicate using them for this purpose generates misleading estimates. In the case of standardized achievement tests, using them to assess teacher effectiveness is not valid for several reasons; the main concern is the fact that assessing teacher effectiveness is not what they were designed to measure (Popham, 1999).

Three different guidelines of the Code of Fair Testing Practices in Education (2005), under the heading Reporting and Interpreting Results, make it clear that tests should not be used for purposes other than those they were designed for, and yet standardized test frequently are used to hold teachers accountable for student learning. Three major reasons that the misuse of test results for teacher effectiveness is erroneous are the alignment of test items and curriculum, adequate opportunity to demonstrate knowledge on the assessment, and problems with separating educational influences from confounding factors.

There are a number of reasons tests design does not provide alignment with curriculum. First, in order for test makers to make their product marketable, they must make the content of the test as broad as possible. The broadness of assessments means that anywhere from 50 to 80% of the test items do not correlate with what is taught (Damore, 2005). In fact, 44% of states have standardized assessments that do not align with their adopted curriculum (Barton, 2005). A second reason is that to create a spread that will sort students into proficiency levels, test makers exclude many test items that deal with content that teachers stressed because of its importance. A final reason is that achievement tests are as likely to measure what has happened in early childhood, home life, and after school as it is to measure what happens in school (Barton, 2004).

Two other areas of validity that relate to test quality that researchers indicate make standardized assessments poor choices for assessing teacher effectiveness are readability and construct-irrelevant variance. Item readability refers to items that are written using above grade-level vocabulary that cause students to miss items due to reading difficulty not because of a lack of content knowledge (Hewitt & Homan, 2004). Construct-irrelevant variance includes readability as well as test preparation that inflates test scores, test item format, adverse testing conditions, and accuracy of passing scores.

The validity issues presented here render standardized test scores inappropriate as a single measure of teacher effectiveness. According to Kane (2002), using test scores to make a decision about educational quality requires evidence of the appropriateness of that use. As the previous details demonstrate, the uncertainty of measuring academic achievement argues against using standardized tests as a signal of teacher effectiveness.

Another limitation of using standardized test scores, as presented in chapter II, is that measuring a teacher's effectiveness using a single set of test scores ignores the influence of test-taking strategies, motivation, and attitude toward test taking. Likewise, it overlooks the fact that it is impossible to differentiate between instructional practices that promote learning and those that narrowly teach to the test (Kupermintz, Shepard, & Linn, 2001). Similarly, Shavelson, Webb, and Burstein (1986) have argued that standardized tests are strictly summative and, therefore, do not represent adequately student cognitive growth or the effect a teacher had on that growth.

Missing Data Regarding Possible Moderator Variables

Another limitation of this study was the inability to analyze important moderator variables. As detailed in the multilevel education model of factors influencing student achievement given in chapter II, the level of achievement that students attain is the result of many factors and not just teacher effectiveness. Such factors (student gender, ethnicity, English Language Proficiency level, and socioeconomic status) were coded but could not be analyzed because so few studies included information that could be used for the analysis or did not include them in their research. Ethnicity and years of teaching were coded to analyze teacher factors. Context factors, important influences on student achievement, were not coded because a review of a few sample studies indicated that this was not an area addressed by studies on the effect of NBC. The small number of variables that could be analyzed served to narrow the focus of the study, limiting the recommendations that can be made in terms of using test scores to evaluate teacher effectiveness. Additionally, their absence decreases the trustworthiness of the findings both of the primary studies and of this meta-analysis (Wayne & Youngs, 2003).

Constraints of Value-added Methods

Disentangling the influence of teachers from the influence of the factor presented in previous subsections purportedly can be accomplished using value-added methods. Seven of the studies excluded from the master list presented in chapter III used value-added methods to assess teacher effectiveness. These studies were excluded because they did not provide the descriptive statistics needed to calculate effect sizes. Rosenthal and DiMatteo (2001) considered this absence in the information provided as a bias in research sophistication. Had these studies provided the necessary descriptive statistics, a comparison between their results and those based on mean proficiency may have yielded a clearer understanding of the discrepant findings in the literature on NBC and student achievement.

Sanders (Sanders & Horn, 1994), the author of the Tennessee Value Added Assessment System (TVAAS) presented in chapter II, asserted that value-added analysis controls for the moderator variables also presented in chapter II and the previous subsection by measuring achievement using gains over time or by including prior achievement as an explanatory variable in a regression equation. When student academic achievement differs because of factors other than teacher effectiveness, however, value-added estimates may be biased by the nonrandom placement of students and teacher (Ballou, 2008). How plausible this bias is remains to be assessed because the value-added system is copyrighted, and the owner will not share documentation that would allow researchers to attempt to replicate findings. Had these studies been included, it would have increased the statistical power of the meta-analysis.

Discussion of Findings

The present meta-analysis was designed to answer three questions that are presented independently in the following sections. The results presented in the previous chapter are addressed according to research question in order to organize the interpretation of the data analysis.

Research Question 1

> *What is the effect on student achievement for students taught by teachers with National Board Certification (NBC) when compared with students taught by teachers who are not board certified?*

Although an examination of the descriptive statistics revealed indicators that, on average, students of NBCTs out perform students whose teachers are not board certified in both reading and mathematics, there also were indicators that this difference in student outcomes is not statistically or practically significant for either reading or mathematics achievement. Additionally, the mean effect size for both reading and mathematics were nearly zero, .09 for reading and .08 for mathematics, demonstrating that sometimes NBCTs were shown to be more effective teachers in terms of student outcomes and sometimes nonNBCTs were more effective.

This meta-analysis was undertaken to generate new evidence by examining the aggregated research findings of empirical studies that explored the relationship between NBC status and student achievement because a review of the literature revealed conflicting evidence regarding the effectiveness of NBCTs in producing high student achievement when compared with nonNBCTs. The inability to find statistically significant and practical data that are suggestive of greater academic achievement on standardized tests by students of NBCTs reflect the fact that there are two distinctively different sets of results amid the research on the NBPTS as an identification of teacher effectiveness. On the one hand, several studies provide evidence of NBCTs having a positive influence on student achievement (Cavalluzzo, 2004; Clotfelter et al., 2006; Goldhaber & Anthony, 2005; Harris & Sass, 2006). On the other hand, there are studies that found negative effects when the students of NBCTs and nonNBCTs are compared using standardized test scores (Benigno, 2005; Sanders et al., 2005).

Cavallluzza’s research includes a data set linking 108,000 high-school student records to their subject area teachers. Cavalluzza found that teachers with NBC had students who made the greatest academic gains. Clotfelter et al. (2006) studied the influence of NBCTs and nonNBCTs on student achievement for fifth grade in reading and mathematics. Using various statistical models, they found that NBCTs were more effective in raising reading achievement scores but not in mathematics. Using education production function to compare the achievement results of third through fifth grade students of NBCTs and other teachers, Goldhaber and Anthony (2005) also found that NBCTs were more effective in teaching reading but not mathematics. Using the same statistical model, Harris and Sass (2006) found that students of NBCTs in third through tenth grade performed better on a test of reading achievement than students of other teachers but not on mathematics achievement tests.

Any summary provided in a meta-analysis of a body of research is only as reliable as the methods used to estimate the effect in each of the primary studies (Garg, Hacka, &Tonelli, 2008). Therefore, the limitations that studies comparing NBCTs and nonNBCTs by Cavalluzzo, (2004), Clotfelter et al. (2006), Goldhaber and Anthony (2005), and Harris and Sass (2005) encountered plague the current study as well. Because these studies did not account for the fact that student test scores are nested within classes within schools, it is likely that their results would have been different, if this clustering had been accounted for (Hakel, Koenig, & Elliott, 2007). Because the current meta-analysis used these primary studies to calculate effect sizes and because the mean effect size for reading was close to zero (.09) and for mathematics was .08, it is likely that the lack of clustering in the primary studies included in this research contributed to results that were not statistically significant.

In comparing the achievement scores of fifth through eighth grade students who had teachers who either were NBC or who were not NBC, Sanders et al. found few statistically significant effects for NBCTs. Similarly, Benigno (2005), studying third through eighth graders found that students of nonNBCTs out performed the students of NBCTs on standardized tests of achievement. The results of Sanders et al. (2005) and the dissertations (Benigno, 2005; Childs, 2006; Falaney, 2006; Fisher, 2005; Rouse, 2004; Silver, 2007; Stephens, 2003) in this study were limited by sample size, as was this meta-analysis. Cohen and Becker (2003) demonstrated the relationship between sample size and statistical power, which is an important cause of variability in this meta-analysis.

One purpose of a meta-analysis is to gain greater accuracy and statistical power by taking advantage of the large sample size resulting from the accumulation of results over multiple studies (Lipsey & Wilson, 2001). Jamie DeCoster (2004) in *Meta-analysis Notes* recommended that at least 30 studies are needed to provide statistical power. Unfortunately, the lack of provision of descriptive statistics by some researchers and the fact that other empirical studies did not look at the same theoretical construct, the current study could only analyze 12 studies which, as Lipsey and Wilson pointed out, may have resulted in finding relationships of meaningful magnitude that are not statistically significant due to low statistical power.

Research Question 2

What is the difference in the effect size of reading and mathematics assessments for students taught by NBCTs when compared with students taught by nonNBCTs?

Based on overlapping confidence intervals, on average, teachers who have NBC have a greater effect on student achievement than teachers who do not have NBC in both reading and mathematics. This difference, however, is neither statistically nor practically significant using Hedges and Hedberg’s (2007) benchmark for effect size magnitudes in educational research.

That reading and mathematics would produce the similar effect sizes was surprising, given that most studies find that NBC has a greater impact on student achievement in reading (Clotfelter et al., 2007; Fisher, 2005; Goldhaber & Anthony, 2005; Harris & Sass, 2007). The difference in findings may be the result of the inclusion of more studies examining the effect of mathematics in the current meta-analysis. Only 9 of the 12 studies examine reading achievement scores, whereas all 12 examined mathematics achievement scores.

Research Question 3

> *To what extent do variables, such as type of assessment, type of publication, and study quality moderate the relationship between Certification Status and academic achievement?*

Six total categories of moderator variables were investigated in the current study using the analog to ANOVA, yielding evidence that none of the categories mediated effect sizes for either reading or mathematics achievement. As presented earlier, although there is no fixed minimum number of studies required for meta-analysis, if the number of studies is too small, the resulting effect size can be unstable, and vary depending on which studies are included (Caird, Scialfa, Ho, & Smiley, 2004). Therefore, the small number of studies included in this analysis is likely to have contributed to a lack of power in investigating the magnitude of effect the variables had on mediating the results of primary studies. Other factors that may have adversely affected the analysis of moderator variables in this meta analysis are presented in the subsections labeled by the six categories that were investigated.

School Level

Rothstein and McDaniel (1989) pointed out that both small numbers of studies and studies with small sample sizes may not be powerful enough to detect small- to medium-sized moderator effects. This may have been the case for the school-level moderator variable category given that 64% of the studies included in this meta-analysis explored the relationship of certification status and student achievement at the elementary level. Very few researchers have used middle- and high-school assessment data to explore the effectiveness of National Board Certification. Therefore, only 22% of the included studies were high-school level, and 14% were of middle-school level. One reason for this difference in grade levels that are found in studies is that almost all high school and many middle schools do not assess reading on standardized achievement tests. Instead they assess subjects taught in high-school English classes including literature, poetry, grammar, vocabulary, and writing (Fisher, 2005).

Assessment Type

As reported in chapter II, criterion-referenced tests are intended to measure how well a person has learned a specific body of knowledge and skills, whereas norm-referenced tests are developed to compare test takers with each other. Because norm-referenced tests are designed to produce great variance in scores (Popham, 1975), the finding that differences between criterion-referenced and norm-referenced tests did not account for more of the variability in this meta-analysis was surprising. Studies like that of Harris and Sass (2007), who examined the influence of NBC using assessment data from the entire state of Florida, found differences in outcomes. Because the state gives both norm-referenced and criterion-referenced tests, Harris and Sass were able to compare the results. The results of the study revealed that the estimates of the effect of NBC were negative for both reading and mathematics using the norm-referenced test, whereas the estimates for the criterion-referenced assessments were positive for both subjects. The difference in Harris and Sass' results between norm-referenced and criterion-referenced assessments may have been the result of the different purposes of the assessments. As with other categories, the finding that assessment type did not account for any of the variance in study results may be a consequence of the small number of studies that could be included in this meta-analysis so further research is need to corroborate the findings of Harris and Sass.

Group Assignment

Eight primary studies in this meta-analysis matched NBCTs and nonNBCTs on teacher effects, including gender, years of experience, and advance degrees. As mentioned in chapter II of this meta-analysis, randomization is not feasible in most school settings for a variety of reasons (Braun, 2005; Dimitrov & Rumrill, 2003; Goe et al., 2008), and, therefore, both blocking and matching frequently are used to control for biased estimations that may result from nonrandom sampling (Stuart & Rubin, 2008). Klar and Donner (1997) have pointed out that there are difficulties in estimating the design effect from pairs design. The difficulties largely arise because there is inherent variation in response between clusters in a matched pair, which totally confounds the effect of an intervention. This confounding implies that such variation cannot be used to obtain a valid estimate of the magnitude of effect from matching and blocking in primary studies that is most likely the reason that the category of group assignment did not account for variability in research results.

Assessment

Several confounding factors may have contributed to the finding that the category of assessment, which explored variability between studies that did and did not use pretests, did not account for variability in research results. First, as with the school-level category, the results of the analog to ANOVA for pretests were most likely affected by the numbers of studies Retrieved used in the current study. Next, as with the group assignment category, in order to properly isolate and nullify confounding variables, students must be assigned randomly to teachers, which is not feasible in most schools and school districts. Other confounding factors that may have contributed to the variability of effect sizes in studies using pretest were presented in chapter II and include families choosing schools and districts and principals placing teachers preferentially (Braun, 2005; Kane & Staiger, 2008b).

Study Methodology

There is very little agreement among researchers regarding what constitutes methodological quality; however, it is clear that the quality of a study affects the results of a meta-analysis (Lipsey & Wilson, 2001). Therefore, methodological variation among studies was investigated as part of this meta-analysis. Study methodologies were collapsed into two groups of similar statistical analyses because of small numbers. Neither the category level with t test, analysis of variance (ANOVA), analysis of covariance (ANCOVA), and multivariate analysis of variance (MANOVA) nor the category level with hierarchical linear (HLM), ordinary least squares (OLS), and regression models mediated test results. This result was surprising given that fact that McCaffrey and Rivkin (2007) found that different statistical models inevitability produce different estimates due to finite sampling error and different controls for confounding.

Publication Type

Publication type did not account for variability in research findings. Only published research and dissertations, however, were located and included in the analysis. The issue of publication irretrievability was discussed in chapter III regarding *fail-safe N***.** The results of the analog to ANOVA for this category may have been influenced by the availability of published articles that produce statistically significant findings. Although unpublished dissertations were included in this meta-analysis, unpublished studies with smaller samples and lower statistical power were not located. To evaluate the effect of not including unpublished works, Rosenthal's *fail-safe N* statistic was calculated. *Fail-safe N* estimates the number of unpublished or unretrieved studies showing a zero effect that would be needed in order to make the effects of the meta-analysis statistically nonsignificant (Long, 2001). The results of the analysis indicated that approximately 9 null-result studies would be required to reduce the combined effect size to a statistically nonsignificant level. Because the reverse was true and the publication category was not indicated as a mediator of study outcomes, it is possible that 9 more studies may have been needed to analyze the category adequately. Given that a literature search of four major data bases produced only 30 studies that assessed the effect of NBC on student achievement scores, it seems unlikely that this many studies exist in researchers' file drawers.

Recommendations for Research

Based on the findings and conclusions of this study, recommendations for future research are indicated. The recommendations are presented in three subsections: (a) *Meta-analysis Sample Size,* (b) *Assessment of Student Achievement* (c) *National Board Certification and Student Achievement.*

Meta-analysis Sample Size

The argument was presented earlier in this chapter that the relatively low number of studies included in the meta-analysis played a role in the outcomes of the analysis of variances conducted to answer *Question 3.* To address the issue of sample size in a meta-analysis, Jamie DeCoster (2004) in *Meta-analysis Notes* recommended that at least 30 studies are needed to provide statistical power. This number may be obtainable for future meta-analysis research as the body of published, unpublished, and dissertations grow in response to policy makers and other stakeholders questioning the expense of NBC.

Given that the societal costs of the NBC model of professional development include the costs of pursuing certification, which averages annually between $18,000 and $31,000 per person and over 400 uncompensated man hours to complete a portfolio (Cohen & Rice, 2005), there will continue to be a call for research that validates the cost effectiveness of the model. As a result of both the scope and expense of board certification, it is critical to continue to study the relationship between board certification and student achievement to help determine its cost effectiveness. The synthesis of the findings of past studies via meta-analysis is important because evidence of teacher effectiveness obtained by assessing student gains adds to the body of research demonstrating the value of the NBC process and, therefore, indicating a positive return on individual and societal investments.

Assessment of Student Achievement

The argument earlier in this chapter under the subheading of *Assessment Types* that the use of criterion-referenced and norm-referenced assessments may account for variability in research outcomes needs to be considered and explored by future researchers. Although this meta-analysis did not find that assessment type mediated research findings, Harris and Sass's (2007) study indicated the opposite. Their research revealed that the estimates of the effect of NBC were negative for both reading and mathematics using the norm-referenced test, whereas the estimates for the criterion-referenced assessments were positive for both subjects. The difference in Harris and Sass' results between norm-referenced and criterion-referenced assessments may have been the result of the different purposes of the assessments presented in chapter II.

Educational Testing Service (ETS) has published guidelines as have other testing organizations that their tests should only be used for the purpose for which they were designed for; however, schools, states, and the Federal Government continue to use them to measure teacher effectiveness. These guidelines echo Shavelson et al.'s (1986) concerns discussed in chapter II. They found three main concerns with using standardized tests to measure student achievement and teacher effectiveness. As mentioned previously, tests are not being used for their intended purposes. According to Shavelson et al., it is questionable to use norm-referenced tests to make inferences about the effectiveness of schools, districts, and, in particular, teachers because they strictly are summative and not adequate measures of cognitive growth. Similarly, grade-level criterion-referenced assessments should not be used because they are linked to grade-level standards and are not sensitive to cognitive growth. Therefore, research into how assessment types influence study outcomes is indicated.

Given the issues surrounding standardized testing, it may be appropriate to conduct mixed method studies of NBC as a signal of teacher effectiveness. In the same manner that process-product research, as presented in chapter II, investigated teacher factors and how they lead to teacher effectiveness, future research could combine the theoretical frameworks of Bond, Smith, Baker, and Hattie (2000) with Ding and Sherman (2006) in a study to obtain a more valid picture of teacher inputs and student outcomes. Accountability data may be more useful if augmented with other sources of information that provide evidence of the effects of educational practice (Raudenbush, 2004).

There are two other areas related to assessment that should be explored further by researchers. One is the fact that curriculum and standardized tests rarely align, which ignores the student variables such as test-taking strategies, motivation, and attitude as well as teacher factors such as instructional practices and teaching to the test (Shavelson et al., 1986). Another concern to research is the use of a single score to assess student achievement and teacher effectiveness, which assumes that all or nearly all of what a student learns in a year, is the result of a single teacher's efforts in a single year.

National Board Certification and Student Achievement

The limited amount of detail in the published research and dissertations made it difficult to analyze critically the research on NBC and the comparison of student achievement scores for NBCTs and nonNBCTs. Therefore, like the previous reviews of studies comparing student of achievement outcomes for students of NBCTs and NBCTs (Hakel, Koenig, & Elliott, 2008; Holland, 2006; Leef, 2003; Predrosky, 2001; Stone, 2002), this meta-analysis failed to answer the question of whether or not National Board Certification identifies effective teachers who increase student learning outcomes. Consequently, because NBC is such an involved process, further research is necessary in order confirm that NBC is a reliable measure of effectiveness for increasing student achievement in order to corroborate a positive return on a candidate's investments of time and energy.

Future studies seeking to investigate the effectiveness of National Board Certification need to increase their level of detail and include all possible variables that can influence student achievement. The consequences of omitting variables may cause discrepant substantial upward bias in the magnitude of effects (Palardy, 2010). Additionally, omitting variables restricts the inferences that can be drawn about the effectiveness of different teachers. Important items to include that can account for unmeasured teacher, student, and contextual factors that contribute to the heterogeneity of findings regarding NBC are shown in Figure 1 of chapter II.

More specifically, teacher factors to include are licensure, measures used to grant licensure, years of experience, advance degrees, verbal ability, and personal traits such as race and gender (Clotfelter et al., 2007; Harris & Sass, 2009; Munoz & Chang, 2007; Wayne & Youngs, 2003). Student factors are race and gender as well as socioeconomic status (Coleman, 1966; Clotfelter, 2007; Noguera, 2008; Stewart, 2007). Contextual factors that future studies need to include are class size, peer, school, neighborhood, and community factors as well as family and background factors (Card & Rothstein, 2007; Darling-Hammond, 2000; Hanushek, Kain, Markman, & Rivkin, 2002; Hattie, 2003; Jargowski & El Komi, 2009). Additionally, when planning value added studies that compare effect of NBCTs with nonNBCTs, methodologists need to consider bias in research sophistication. Rosenthal and DiMatteo (2001) described sophistication bias studies, like the primary studies included in this meta-analysis, as those that do not take into account sufficient teacher, student, and contextual characteristics.

Finally, a synthesis of the result of the value-added studies excluded from this meta-analysis is needed to compare the differences and similarities with the findings of the current meta-analysis. Value-added criteria are becoming more frequent in assessing teacher effectiveness; however, there continues to be concerns that value-added methods do not permit comparisons of teachers across schools, which would be necessary to evaluate the effectiveness of NBC (Ballou, 2005b). Also of concern is that value-added connotes a causal relationship because it evaluates how a teacher adds value to what a student already knows (Raudenbush, 2004), thus, making value-added appear more valid than the accountability systems that use school mean achievement. The greater validity of value-added studies remains to be tested; therefore, researchers need to conduct additional studies to learn if the causal relationship is defensible. Raudenbush (2004) did conduct one such study and found that both methods, those based on mean proficiency and those using value-added measures, had considerable uncertainty and some unknown bias. If these methods are to be used to investigate teacher effectiveness related to Board Certification, further research needs to be conducted to learn if there are differences in outcomes between the two methods.

Conclusion

As evidenced by the literature in chapter II, research on effective teaching is an important and highly evaluated area of education. Although research into the effectiveness of NBPTS is a relatively new field, it reflects the value placed on studying teacher factors that influence student achievement. Notwithstanding the fact that the results of research on the effectiveness of NBC on student achievement continues to be inconclusive, there are sufficient, albeit small, statistically significant results in the current study to warrant the continued investigation of National Board Certification's influence on student achievement. Therefore, more data are needed to generate useable descriptive statistics to further examine the tenuous findings in the literature that NBC is related to student achievement scores.

Likewise, more data are needed to probe the differences between NBCT's influence on reading and mathematics achievement. Although the results of this meta-analysis indicated that there is no difference between achievement outcomes for students of NBCTs in reading and mathematics, most studies report that NBCTs have the greatest impact on student achievement in mathematics. Further investigating the factors that contribute to differences in outcomes in mathematics and reading would be worthwhile in order to determine if there are some additional criteria that could be used to better prepare teachers of reading.

Finally, there is a need for the body of research to improve its sophistication by including all variables that that a review of the literature determined may mediate findings regarding student achievement. Further research would be beneficial in understanding more deeply the influence of teacher, student, and contextual factors on student achievement.

References

References marked with an asterisk indicate both included and excluded studies for the meta-analysis.

Alexander, C. (2004, April). *Does teacher certification matter?: Teacher certification and middle school mathematics achievement in Texas*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Aloe, A.M., & Becker, B.J. (2009). Teacher verbal ability and school outcomes: Where is the evidence? *Educational Researcher, 38*(8), 612-623.

American Psychological Association (2002). *Publication manual of the American Psychological Association* (5th ed.), Washington DC: Author.

American Research Council. (1999). *Executive summary: To touch the future; transforming the way teachers are taught.* Washington, DC. Retrieved February 18, 2010, from http://www.acenet.edu/bookstore/pdf/teacher-ed-rpt.pdf

Amrein-Beardsley, A. (2008). Methodological concerns about the Educational Value-added Assessment System. *Educational Researcher, 37*(2), 65-75.

Amrein-Beardsley, A. (2009). Value-added tests: Buyer be aware. *Educational Leadership, 67*(3), 38-42.

*Angle, J.M. (2006). *Science teacher efficacy, national board certification, and other teacher variables as predictors of Oklahoma students' end-of-instruction (EOI) biology I test scores*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3211667).

Ariel, B. & Farrington, D. (2010). Randomized block designs. *The Annals of the American Academy of Political and Social Science, 71*(1), 578-595.

Ballou, D. (2005a). Estimating teacher quality from student test scores. *The Quarterly Journal of Economics, 111*(1), 97-133.

Ballou, D. (2005b, May). *Value added assessment: Controlling for context and misspecified models*. Paper presented at the Third Research Seminar in Analytic Issues in the Assessment of Student Achievement, Washington, DC.

Ballou, D. (2008, November). *Value-added analysis: Issues in the economic literature*. Paper presented at the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC. Retrieved February 18, 2010, from http://www7.nationalacademies.org /BOTA/VAM%20Analysis%20-%20Ballou.pdf

Ballou, D., Sanders, W., & Wright, P.S. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37-66.

Barton, P.E. (2004). Unfinished business: More measured approaches in standards-based reform. *Educational Testing Service Policy Brief*. Retrieved February 18, 2010, from http://www.ets.org/Media/Education_Topics/pdf/unfinbusiness.pdf

*Benigno, S.C. (2005). *A comparison of student scores on the Mississippi curriculum test of students taught by National Board certified teachers and non-National Board certified teachers*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3209666).

Bell, C.A., Little, O.M., Croft, A.J., & Gitomer, D.H. (2009, April). *Measuring teaching practice: A Conceptual Review*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Berliner, D. (2009). Are teachers responsible for low achievement by poor students? *Kappa Delta Pi Record*, *46*(1), 18-22.

Bezzina, M. (1986). The contribution of process-product-research to the identification of effective teaching skills. *Asia-Pacific Journal of Teacher Education*, *14*(2), 38-44.

Blank, R.K. (2004). *Data on enacted curriculum study: Summary of findings*. Council of Chief State School Officers, Washington, DC. Retrieved May 5, 2010, from http://seconline.wceruw.org/Reference/DECStudy04.pdf

Blanton, L., Sindelar, P.T., & Correa, V. (2006). Models and measures of beginning teacher quality. *The Journal of Special Education*, *40*(2), 115-127.

Blanton, L., Sindelar, P.T., Correa, V., Hardman, M., McDonnell, J., & Kuhel, K. (2003). *Conceptions of beginning teacher quality: Models for conducting research.* Gainsville, FL: Center on Personnel Studies in Special Education. COPSSE Document No. RS-6.

Bond, L.B., Smith, T., Baker, W. K., & Hattie, J.A. (2000). *The certification system of the National Board For Professional Teaching Standards: A construct and consequential validity study*. Greensboro, NC: University of North Carolina at Greensboro.

Braun, H.I. (2005). *Using student progress to evaluate teachers: A primer on value-Added models.* Retrieved February 18, 2010 from http://www.ets.org/Media/Research/pdf/PICVAM.pdf

*Bundy, J. (2006). *The effect of National Board Certified teachers on average student achievement in North Carolina schools* (Unpublished master's thesis). University of North Carolina at Chapel Hill, Chapel Hill, NC. Retrieved March 10, 2009, from http://www.sog.unc.edu/uncmpa/pdfs/capstone/MPA%20Capstone %20paper%20Bundy%202006.pdf

Burke, M.A., & Sass, T.R,. (2008). Classroom Peer Effects and Student Achievement. Working Paper 08-5. *Federal Reserve Bank of Boston*, Retrieved April 17, 2009, from http://www.caldercenter.org/partner/texas.cfm

Cabezas, C.C. (2006). *The influence of highly qualified teacher designation, and other teacher variables, on student achievement.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3226820).

Card, D., & Rothstein, J. (2007). Racial segregation and the black-white test score gap. .*National Bureau of Economic Research*, Working Paper, 6691. Retrieved February 14, 2010, from http://www.nber.org/papers/w12078.

Caird, J.K., Scialfa, C., Ho, G., & Smiley, A. (2004). The effects of cellular telephones on driving behavior and accident risk: results of a meta-analysis. *Proceedings of the Third International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design.* Retrieved May 5, 2010, from www.nsc.org/...A% 20 meta-analysis%20of%20 driving%20performance%20and%20crash%

Camp, D. (2007). *Where do standards come from? A phenomenological study of the development of National Board Early Childhood/Generalist Standards.* Retrieved March 6, 2009, from http://www.thefreelibrary.com/Where+do+standards+ come+from%3F+A+phenomenological+study+of+the+...-a0167306081

Campbell, R.J., Kyriakides, L., Muijs, R.D., & Robinson, W. (2003). Differential teacher effectiveness: Towards a model for research and teacher appraisal. *Oxford Review of Education*, *29*(3), 347-362.

*Cantrell, S., Fullerton, J., Kane, T.J., & Douglas Staiger, D.O. (2007). *National Board Certification and teacher effectiveness: Evidence from a random assignment experiment.* Retrieved March 24, 2009, from http://www.gse.harvard. edu/~ pfpie/pdf/ National_ Board_Certification.pdf

Caro, D.H. (2009). Socio-economic status and academic achievement trajectories from childhood to adolescence. *Canadian Journal of Education.* Retrieved March 2, 2010, from http://www.highbeam.com/doc/1P3-1885056351.html

*Cavalluzza, L. (2004). *Is National Board Certification an effective signal of teacher quality?* Retrieved February 24, 2006, from http://www.nbpts.org/UserFiles/File/ Final_Study_11204_D_-_Cavalluzzo_-_CNA_Corp..pdf

*Childs, D.E. (2006). *Elementary school National certified teachers and student Achievement* (Doctoral dissertation). Retrieved from ProQuest

Dissertations and Theses database. (UMI No. 32224419).

Clotfelter, C. T., Ladd, H.F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness, *Journal of Human Resources, 11*(4), 778-820.

*Clotfelter, C. T., Ladd, H.F., & Vigdor, J. L. (2007). *How and why do teacher credentials matter for student achievement?* CALDER Working Paper, 2. Retrieved March 24, 2009, from http://www.caldercenter.org/pdf/1001058_teacher_credentials.pdf

Cohen, C.E., & Rice, J.K. (2005). National Board Certification as Professional Development: Design and Cost. *U.S. Department of Education and the National Science Foundation.* Washington, DC. Retrieved March 24, 2009, from http://www.nbpts.org/UserFiles/File/Complete_Study_Cohen.pdf

Cohen, L.D., & Becker, B.J. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8*(3), 243-253.

Cooper, H. M. (1998). *Synthesizing Research: A Guide for Literature Reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Cooper, H., & Hedges, L.V. (1993). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Cornell Statistical Unit. (2008). Overlapping Confidence Intervals and Statistical Significance.*StatNews#3.* Retrieved May 4, 2010, from *http://www.cscu. cornell.edu/news/statnews/ stnews73.pdf*

Coleman, J.S. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

Cooper, H. (1998). *Synthesizing research,* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Cumming, J., & Maxwell, G. (1999). Contextualizing authentic assessment. *Assessment in Education*, *6*(2), 177-194.

Cunningham, G.K., & Stone, J.E. (2005). Value-added assessment of teacher quality as a alternative to National Board for Professional Teaching Standards: What recent studies say. In R. Lissilz (Ed.), *Value added models in education: Theory and applications,* (pp.37-65). Maple Grove, MD: JAM Press.

Damore, S.J. (2005). Why can't you improve my child's test score? *The Charter Schools Resource Journal, 1*(1), 1-89.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Educational Policy Analysis Archives, 8*(1), 1-24.

Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record*, *106*(6), 1047-1085.

Darling-Hammond, L., & Youngs, P. (2002). Defining “high quality teachers”: What does scientifically based research actually tell us?. *Educational Researchers, 31*(9), 13-25.

DeCoster, J. (2004). *Meta-analysis notes*. Retrieved December 28, 2009, from http://www.stat-help.com/notes.html

de Liz, T.M., & Stauss, B. (2005). Differential efficacy of group and individual/couple psychotherapy with infertile patients. *Human Reproduction*, *20*(5), 1324-1332.

Ding, C., & Sherman, H. (2006). Teaching effectiveness and student achievement: Examining the relationship. *Educational Research Quarterly, 29*(4), 39-49.

Dimitrov, D., & Rumrill, P. (2003). Speaking of research. Pretest-posttest designs and measurement of change. *Work*, *20*(2), 159-165.

Doyle, W. (2010). Paradigms for research. on teacher effectiveness. *Review of Research in Education*, *77*(5), 163-198.

Durlak, J.A. (1995). Understanding meta-analysis. In L.G. Grimm & P.R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (6th ed., pp. 319-352). Washington, DC: American Psychological Association.

Field, A. (2005). *Discovering statistics using SPSS*. Thousand Oaks, CA: Sage Publications.

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

*Falaney, P.E. (2006). *National Board for Professional Teaching Standards certification: Does it affect student learning?* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3257510).

*Fisher, S. (2005). *A study of the relationship between the National Board Certification status of teachers and students' achievement*. Technical Report. Columbia: South Carolina, Department of Education. Retrieved January 12, 2010, from http://www.education-consumers.com/articles/South%20Carolina%20NBPTS%20report.pdf

Francis, B., & Skelton, C. (2005). *Reassessing gender and achievement: Questioning contemporary key debates.* London: Routledge & Kegan Paul.

Galton, M. (1987). An ORACLE chronicle: A decade of classroom research. *Teaching and Teacher Education, 3*(4), 299-313.

Gage, N.L. (1972). *Teacher effectiveness and teacher education: The search for a scientific basis.* Palo Alto, CA: Pacific Books.

Gage, N.L., & Needels, M. (1989). Process-product research on teaching: A review of criticisms. *The Elementary School Journal, 89*(3), 253-300.

Garg, A.X., Hackam, D., & Tonelli, M. (2008). Systematic review and meta-analysis: When one study is just not enough. *American Society of Nephrology, 3,* 253-260.

Glazer, E.M., & Hannafin, M.J. (2006). The collaborative apprenticeship model: Situated professional development within school settings. *Teaching and Teacher Education*, 22, 179-193.

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved February 14, 2010, from http://www. tqsource.org/publications/Link BetweenTQandStudentOutcomes.pdf.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

*Goldhaber, D., & Anthony, E. (2003). *Can teacher quality be effectively assessed? National Board Certification as a signal of effective teaching.* University of Washington and the Urban Institute. Retrieved February 18, 2006, from http://www.urban.org/url.cfm

Goldhaber, D., & Hansen, M. (2008). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions*. CALDER Brief 3. Retrieved April 17, 2009, from http://www.caldercenter.org/ upload/ Teacher_ Job_Performance.pdf

Goldhaber, D., Perry, D., & Antony, E. (2007). The National Board for Professional Teaching Standards (NBPTS) process: Who applies and what factors are associated with NBPTS certification? *Education and Policy Analysis, 26*(4), 259-280.

Good, T.L. (1979). Teacher effectiveness in the elementary school. *Journal of Teacher Education, 30*, 52-64.

Good, T.L., & Grouws, D.A. (1977). Teaching effects: A process-product study in fourth grade mathematics classroom. *Journal of Teacher Education, 28*(3), 49-54.

Gordon, R., Kane, T.J., & Staiger, D.O. (2006). *Identifying effective teachers using performance on the job: The Hamilton Project. Discussion Paper*, 2006-01. Washington, DC: The Brookings Institute.

Gronlund, N.E. (2006). *Assessment of student achievement* (8th ed.). Boston, MA: Pearson.

Hakel, M.D., Koenig, J.A., & Elliott, S.W. (2008). *Assessing accomplished teaching: Advanced level certification programs*. Washington, DC: The National Academies Press.

Haladyna, T.M., & Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Hanushek, E.A. (2004). Teacher Quality. In L.C. Solomon, & T.W. Schiff (Eds.), *Teacher Quality &Talented Teachers: The Essential Force for Improving Student Achievement* (pp. 1-12). Santa Monica, CA: Milken Family Foundation.

Hanushek, E.A., Rifkin, S.G., & Kain, J.F. (1998). *Teachers, schools and academic achievement.* National Bureau of Economic Research, Working Paper 6691. Retrieved February 14, 2010, from http://www.nber.org/papers/w6691.

Hanushek, E.A., Kain, J.F., Markman, J.M., & Rivkin, S.G. (2002). *Does peer ability affect student achievement?* National Bureau of Economic Research. Working Paper 14607. Retrieved February 14, 2010, from http://opr.princeton.edu/ seminars/papers/kane_and_staiger3-30- 2010.pdf

Harris, D. (2005, June). *Toward an economics-based theory of educational accountability*. Paper presented at the conference of the Institute for Research on Economics and Sociology of Education, Dijon, France.

Harris, D. (2008). Would accountability based on teacher value-added be smart policy?: An examination of the statistical properties and policy alternatives. *Education Finance and Policy, 4*(4), 319-350.

*Harris, D.N., & Sass, T.R. (2007). *The effects of NBPTS-certified teachers on student achievement.* CALDER Working Paper, 4. Retrieved February 6, 2010, from http://www.caldercenter.org/PDF/1001058_Teacher_Credentials.pdf

Harris, D.N., & Sass, T.R. (2009). *What makes for a good teacher and who can tell?* CALDER Working Paper 30. Retrieved April 17, 2009, from http://www. caldercenter.org/partners/texas.cfm

Hattie, J. (2003, October). *Teachers make a difference: What is the research evidence?* Paper presented at the conference of the Australian Council for Educational Research on: Building Teacher Quality, Melbourne, Australia.

Harwell, M., & LeBeau, B. (2010). Student eligibility for a free lunch as an SES measure in educational research. *Educational Researcher, 39*(2), 120-131.

Hay, I., Ashman, A.F., & Van Kraayenoord, C.E. (1998). The influence of gender, academic achievement and non-school factors upon pre-adolescent self-concept. *Educational Psychology*, *18*(4), 461-471.

Haycok, K. (1998). *Good teaching matters: How well-qualified teachers can close the gap*. Education Trust, Washington, DC. Retrieved February 14, 2010, from http://www.edtrust.org/dc/publication/good-teaching- matters-how-well-qualified-teachers-can-close-the-gap

Hedges, L.V., & Hedberg, E.C. (2007). Interclass correlation values for planning-randomized trials in education. *Education Evaluation and Policy Analysis*, *29*(1), 60-87.

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedwig, L. (2007). The effects of school racial and ethnic composition on academic achievement during adolescence. *The Journal of Negro Education, 76*(2), 154-172.

Heneman III, H., Milanowski, A., Kimball, S., & Odden, A. (2006). Standards-based teacher evaluation as a foundation for knowledge-and skill-based pay. *CPRE Policy Briefs.* Retrieved February 13, 2008, from www.cpre.org

Hewitt, M.A., & Homan, S. P. (2004). Readability level of standardized test items and student performance: The forgotten validity variable. *Reading Research and Instruction, 43*(2), 1-16.

Hill, C.W. (1921). The efficiency ratings of teachers. *The Elementary School Journal, 21*, 438-443.

Hill, P.K. (2005). *The influence of dispositions of effective middle school teachers on teacher behavior: implications for teacher preparation programs*. Symposium on Educator Dispositions. Retrieved February 3, 2010, from http://coehs.nku.edu/ educatordispositions/symposium_2007/papers/InfluenceofDispositions.pdf

Hinkle, D.E., Wiersma, W., & Jurs, S.G. (2003). *Applied statistics for the behavioral Sciences* (5th ed.). Geneva, IL: Houghton Mifflin.

Huang, F.L., & Moon, T.R. (2009). Is experience the best teacher? A multilevel analysis of teacher characteristics and student achievement in low performing schools. *Educational Assessment*, *Evaluation and Accountability, 21*(2), 209-234.

*Holland, J. W. (2006). *Are Mississippi students achieving at a higher rate as a result of National Board Certified teachers?* (Doctoral dissertation). Retrieved from Dissertations & Theses: A&I database. (A&I No. AAT 3238928).

Ingvarson, L., & Hattie J. (2008). *Assessing teachers for professional certification*. Bingley, UK: Elsevier Science Ltd.

Jargowsky, P.A., & El Komi, M. (2009). *Before or after the bell?: School Context and neighborhood effects on student achievement*. CALDER Working Paper 28. Retrieved April 17, 2009, from http://www.caldercenter.org/partners/texas.cfm

Joint Committee on Testing Practices. (2005). Code of Fair Testing Practices in Education. *Educational Measurement: Issues and Practice, 24*(1), 23-26.

Kane, M. (2005). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31-41.

*Kane, T.J., & Staiger, D.O. (2008a). *Are teacher-level value-added estimates biased?: An experimental validation of non-experimental estimates*. Retrieved April 22, 2010, from isites.harvard.edu/fs/docs/icb.topic245006.files/Kane_Staiger_3-1 7-08.pdf.

Kane, T.J., & Staiger, D.O. (2008b*). Estimating teacher affects on student achievement: an experimental evaluation.* NBER. Working Paper, 14607. Retrieved February 14, 2010, from http://opr.princeton.edu/seminars/papers/kane_and_staiger3-30-2010.pdf

Kennedy, M. (2006). From teacher quality to quality teaching. *Educational Leadership*. *63*(2), 14-18.

Klar, N., & Donner A. (1997) The merits of matching in community intervention trials. *Statistics in Medicine, 16*(46), 1753–1764.

Knapp, T. R., & Schafer, W. D. (2009). From gain score t to ANCOVA F (and vice versa). *Practical Assessment, Research & Evaluation*, *14*(6). Retrieved February 14, 2010, http://pareonline.net/getvn.asp?v=14&n=6.

Koedel, C., & Betts, J.R. (2007). *Re-examining the role of teacher quality in the educational production function*. Working Paper #2007-03. Nashville, TN: National Center on Performance Initiatives.

Kupermintz, H. (2003). Teacher effects and effectiveness: A validity investigation of the Tennessee value added assessment system. *Educational Evaluation and Policy Analysis*, *25*(3), 287-298.

Kupermintz, H., Shepard, L., & Linn, R. (2001, April). *Teacher effects as a measure of teacher effectiveness: Construct validity considerations in TVASS (Tennessee Value Added Assessment System)*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle.

Lederman, N.G., & Niess, M.L. (2001). An attempt to anchor our moving targets. *School Science and Mathematics, 101*(2), 57-59.

Leef, G.C. (2003). *National Board Certification: Is North Carolina getting its money's worth?* A Policy Report from the NC Education Alliance. Retrieved April 19, 2009, from www.nceducationalliance.org

Light, R.J., & Pillemer, D.B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Lipsey, M.W., & Wilson, D.B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Long, J. (2001, February). *An introduction to and generalization of the "fail-safe N."* Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans.

Lustick, D., & Sykes, G. (2006). National board certification as professional development: What are teachers learning? *Education Policy Analysis Archives, 14*(5), 1-46.

Markley, T. (2004). *Defining the effective teacher: Current arguments in education.* Retrieved February 14, 2010, from http://www.usca.edu/essays/vol112004/markey.pdf.

Marshall, J. (2006). Uniting the Five Core Propositions and Effective Teacher Dispositions. *Teacher Education Journal of South Carolina, 8*(1), 43-46.

McCaffrey, D.F., Koretz, D.L., Lockwood, J.R., & Mihally, K. (2004). *Evaluating Value- added models for teacher accountability.* Santa Monica, CA: RAND Corporation.

McCaffrey, D.F., & Rivkin, S.G. (2007*). Empirical investigations of the effects of national board teacher standards certified teachers on student outcomes.* Paper prepared for the Committee on Evaluation of Teacher Certification by the National Board for Professional Teaching Standards. Retrieved November 17, 2009, from, http://www7.nationalacademies.org/bota/NBPTS- MTG4-

McCaffrey-Paper.pdf

McCaffrey, D.F., Sass, T., & Lockwood, J.R. (2009). The intertemporal stability of teacher effects. *Education Finance and Policy, 4*(4), 572-606.

*McColskey, W., Stronge, J. H., Ward, T. J., Tucker, P. D., Howard, B., Lewis, K., & Hindman, J.L. (2005). *Teacher effectiveness, student achievement, and National Board Certified teachers*. Arlington, VA: National Board of Professional Teaching Standards.

Medley, D.M. (1972a). Early history of research on teacher behavior. *International Review of Education, 18*(4), 430-439.

Medley, D.M. (1972b). *Teacher competence and teacher effectiveness: A review of process-product research*. Washington, DC: American Association of Colleges for Teachers.

Munoz, M.A., & Chang, F.C. (2007). The elusive relationship between teacher characteristics and student academic growth: A longitudinal multilevel model for change. *Journal of Personnel Evaluation in Education, 20*, 147-164.

National Assessment of Education Progress. (2009). *Nations report card.* Retrieved February 14, 2010, from http://nces.ed.gov/nationsreportcard/

National Board for Professional Teaching. (1989). *What teachers should know and be able to do. An NBPT Report*. Arlington, VA: Author.

National Board for Professional Teaching. (2001). *Policy for networks of National Board Certified teachers and friends.* Retrieved February 14, 2010, from http://www.nbpts.org/UserFiles/File/networkpolicy.pdf

National Board for Professional Teaching Standards. (2002). *Standards development.* Retrieved January 28, 2010, from www.nbpts.org/the_standards/standards developmental.

National Board for Professional Teaching Standards. (2005). *What teachers should know and be able to do: The five core propositions of the National Board*. Retrieved June 25, 2005, from http://www.nbpts.org/about/coreprops.cfm

National Commission on Excellence in Education. (1983). *A Nation at Risk: The imperative for reform: A Report to the Nation and Secretary of Education, United States Department of Education.* Washington, DC: Author.

National Commission on Teaching and America’s Future. (1996). *What matters most: Teaching for America's future: Report of the National Commission on Teaching America's Future.* Woodbridge, VA: Author.

National Commission on Excellence. (2004). *Report to the Nation and the Secretary of Education*, Washington, DC: The Commission [Supt. of Docs., U.S. G.P.O. distributor].

Noguera, P.A. (2008). Creating schools where race does not predict achievement: The role and significance of race in the racial achievement gap. *The Journal of Negro Education, 7*(2), 90-104.

Nye, B., Konstantopoulos, S., & Hedges, L. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237-257.

Odden, A., Borman, G., & Fermanich, M. (2004). Assessing teacher, classroom and school effects, including fiscal effects. *Peabody Journal of Education, 79*(4), 4-32.

Olejnik, S.F. (1988). Variance heterogeneity: An outcome to explain or a nuisance factor to control. *Journal of Experimental Education*, *56*(4), 193-197.

Oliver, S. (2010). *Certification vs. licensure.* Retrieved February 19, 2010, from http://www.cbmt.org/default.asp?page=Certification%20vs.%20Licensure

Palardy, G.J. The multilevel crossed random effects growth model for estimating teacher and school effects: Issues and extensions. *Educational and Psychological Measurement*, *20*(20), 1-19.

Palmer, D.J., Stough, L.M., Burdenski, T.K., & Gonzales, M. (2001, April). *Identifying Teacher expertise: An examination of researchers' decision making*. Paper presented at the annual meeting of the American Educational Research Association, Seattle.

Popham, W.J. (1975). *Educational evaluation.* Englewood Cliffs, NJ: Prentice Hall, Inc.

Popham, W.J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership, 56*(6), 8-15.

Popham, W.J. (2000). Stopping the mismeasurement of educational quality. *The School Administrator*, 5*7*(11), 12-15.

Porter, B.A. (2009). *The effects of being placed in special education classes versus general education classes and teacher certification on students' high-stakes testing scores.* (Doctoral dissertation). Retrieved from Electronic Dissertation Collection Database. (EDC No. etd-08122009-082312).

Predrosky M. (2001). Defrocking the National Board: Will the imprimatur of "board certification" professional teaching. *Education Matter.* Retrieved April 7, 2006, from http://www.edweek.org/ew/articles/2001/0411/30

Rice, J. K., & Hall, L.J. (2008). National Board Certification as professional development: What does it cost and how does it compare? *Education Finance and Policy*, *3*(3), 339-373.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (rev.). Newbury Park, CA: Sage Publications.

Rosenthal, R., & DiMatteo, M.R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*(1), 59-82.

Rothstein, J. (2009). *Teacher quality in educational production: Tracking, decay and student achievement.* Princeton University and NBER. Working Paper, 14442. Retrieved February 14, 2010, from http://www.nber.org/papers/w14442.

Rothstein, H. B., & McDaniel, M.A. (1989). Guideline for conducting and reporting meta-analyses. *Psychological Reports, 65*, 759-770.

*Rouse, W.A. (2004). *An examination of student test results: National Board-Certified teachers and non-National Board-Certified teachers.* (Doctoral dissertation). Retrieved from Dissertations & Theses: A&I database. (A&I No. AAT 3120274).

*Rouse, W., & Hollomon, H.L. (2005). A comparison of student test results: Business and marketing education National Board Certified teachers and non-National Board teachers. *The Delta Phi Epsilon Journal, 47*(3), 128-142.

Rowan, Chiang, B. F., & Miller, R.J., (1997) Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of Education*, *70*, 256-284.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the prospects study of elementary schools. *Teachers College Record*, *104,* 1525-1567.

Rutkowski, E.G., Gonzales, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142-151.

*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.* Arlington, VA: National Board for Professional Teaching Standards.

Sanders, W. L., & Horn, S. (1994). The Tennessee Value-added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299-311.

Sanders, W. L., & Rivers, J.C. (1996). *Cumulative and residual effects of teachers on future academic achievement*. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Sanders, W.L., Saxton, A., Schneider, J., Dearden, B., Wright, S.P., & Horn, S. (2002). *Effects of building change on indicators of student achievement growth: Tennessee Value-Added Accountability System.* Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Sartawi, A.M. (2009). Perceptions of effectiveness in teaching by special and general education teachers in the United Arab Emirates. *International Journal of Disability, Community & Rehabilitation, 8* (3), 17-33.

Seidel, T., & Shavelson, R.J. (2007). Teaching effectiveness in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454-499.

Shavelson, R.J., Webb, N.M., & Burstein, L. (1986). Measurement of teaching. In M. Wittrock (Ed.)*, Handbook of research on teaching (*3rd ed., pp. 50-91). New York: McMillan.

Shulman, L. (1986a). Paradigms and research programs in the study of teaching: A contemporary perspective. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 3-36). New York: McMillan.

Shulman, L. (1986b*).* Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1-22.

Shulman, L. (2000, Fall). Teachers with National Board Certification outperform others in 11 of 13 areas, significantly enhance student achievement, study finds. *The Professional Standard, 1,* 1, 8.

*Silver, K. (2007).*The National Board effect: Does the certification process influence student achievement?* (Doctoral dissertation). Retrieved from Dissertations & Theses: A&I database. (A&I No. AAT 3280759).

Silver, E.A., Mesa, V.M., Morris, K.A., Star, J.R., & Benken, B.M. (2009). Teaching mathematics for understanding: An analysis of lessons submitted by teachers seeking NBPTS certification. *American Educational Research Journal, 46*(2), 501-531.

Smith, T. (2004). Toward a prototype of expertise in teaching: A descriptive case study. *Journal of Teacher Education, 55*(4), 201-213.

Smith, T., Gordon, B., Colby, S., & Wang, J. (2005). *An examination of the relationship between depth of student learning and National Board Certification status.* Arlington, VA: National Board for Professional Teaching Standards.

Stephens, A. D. (2003). The relationship between National Board Certification for teachers and student achievement.* (Doctoral dissertation). Retrieved from Dissertations & Theses: A&I database. (A&I No. AAT 3084814).

Stewart, E.B. (2008). School structural characteristics, student effects, peer associations, and parental involvement: The influence of school- and individual-level factors on academic achievement. *Education and Urban Society, 40*(2), 179-204.

*Stone, J. (2002). *The value-added achievement gains of NBPTS-certified teachers in Tennessee: A brief report.* College of Education, East Tennessee State University. Retrieved January 18, 2005, from http://www.educationconsumers.com/briefs/stoneNBPTS.shtm

Stuart, E.A., & Rubin, D.B. (2008). Matching multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics*, *33*(3), 279-306.

Thalheimer, W., & Cook, S. (2002). How to calculate effect sizes from published research: A simplified methodology. *Work-Learning Research.* Retrieved March 18, 2010, from www.work-learning.com

U.S. State Department of Education, Office of Post Secondary Education. (2004). *The Secretary's fourth annual report on teacher quality*, Washington, DC: Educational Publications Center.

Valentine, J. C., & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes*. Washington, DC: What Works Clearinghouse.

* Vandevoort, L., Amrein-Beardsley, A., & Berliner, D. (2004). National Board Certified teachers and their students' achievement. *Education Policy Analysis Archives, 12*(46), 1-117.

Veldman, D.J., & Brophy, J.E. (1974). Measuring teacher effects on pupil achievement. *Journal of Educational Psychology*, *68*(3), 319-324.

*Vitale, T. (2008) *What is the relationship between National Board Certification and the achievement results of third grade students in a local central Florida school district?* (Doctoral dissertation). Retrieved from Dissertations & Theses: A&I database. (A&I No. AAT 3319281).

Watson, A., Kehler, M., & Martino, W. (2010). The problem of boys' literacy underachievement: Raising some questions. *Journal of Adolescence and Adult*

Literacy, 53(5), 356-361.

Wayne, A.J., & Youngs, P. (2003). Teacher Characteristics and student achievement gains: A review. *Review of Educational Research, 73*(1), 89-122.

Wenglinsky, H. (2004). Facts of critical thinking skills? What NAEP results say. *Educational Leadership, 62(1)*, 32-35.

Weinberg, S.L., & Abramowitz, S.K. (2002). *Data analysis for the behavioral sciences using SPSS*. Cambridge, MA: Harvard University Press.

Wright, S.P., Horn, S.P., & Sanders, W.L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57-67.

Yates, G., Chandler, M., & Westwood, P. (1987). Teacher effectiveness and process-product research: Another look. *The South Pacific Journal of Teacher Education, 15*(2), 18-26.

APPENDIX A

CODING PROTOCOL

STUDY LEVEL CODING MANUAL

Biographic Reference

Write a complete citation in APA form

1. Study ID Number: Each study is assigned a unique ID number. Where a report presents two independent studies with different participants a decimal is added to the study ID number to distinguish the coding for each study within the report separately.

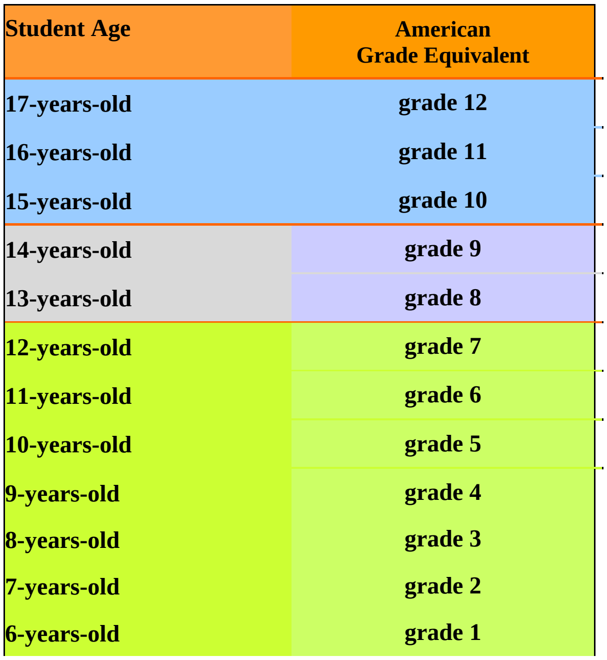
2. Publication ID Number: Each publication is assigned a unique ID number.

3. What type of publication is the report? If two separate reports are being used to code a single study, code the type of the more formally published report (i.e., book or journal).
   1 Peer reviewed
   2 Dissertation
   3 Unpublished Conference paper
   4 Web based
   5 Other:

4. What is the publication year? (last two digits). Where two separate reports are being used to code a single study, code the publication year of the more formally published report.

Sample Descriptors

5. Mean age of the student sample. Specify the approximate or exact mean age at the beginning of the study. Code the best information available; **Use the chart below** to estimate mean age from grade levels when necessary. If mean age can't be determined code "99.99".



Student Age	**American Grade Equivalent**
17-years-old	**grade 12**
16-years-old	**grade 11**
15-years-old	**grade 10**
14-years-old	**grade 9**
13-years-old	**grade 8**
12-years-old	**grade 7**
11-years-old	**grade 6**
10-years-old	**grade 5**
9-years-old	**grade 4**
8-years-old	**grade 3**
7-years-old	**grade 2**
6-years-old	**grade 1**

5-years-old	**Kindergarten**

If available put the average for each grade level.
For example in 6th grade use 11.0 as the mean age, so if the ages three students in 5th, 6th, and 7th are averaged the average (10 + 11+ 12) / 3 = 11.

6. Code the grade levels of student participants. Include the percentage of each group when available.
   1. Elementary (k-5)
   2. Elementary (k-6)
   3. Middle School (6-8)
   4. High School(9-10)
   5. Write in grade levels if different from above list ______________________
      For example if the grade span is K – 3 or 3 – 7 then write the exact grade span included in the study.

7. Racial makeup of student participants. Include the percentage of each group when available.
   1. Caucasian:
   2. African Am
   3. Asian
   4. Hispanic:
   5. Native Am.
   6. Unknown:

8. Gender makeup of student participants. Include the percentage of each group when available.
   1. Male:
   2. Female.
   3. unknown

9. Select the code that best describes the socio-economic status of student participants. Include the criteria used for determining economic status and the percentage of each group when available.
   1. affluent
   2. middle class
   3. poor
   4. unknown

10. Geographic location
    1. Rural
    2. Metropolitan
    3. Urban
    4. Other _________

11. Gender makeup of teacher participants. Include the percentage of each group when available.
    1. Male
    2. Female
    3. unknown

12. Racial makeup of teacher participants. Include the percentage of each group when available.
    1. Caucasian
    4. Hispanic:

2. African Am
3. Asian
5. Native Am.:
6. Unknown

13. Amount of teaching experience for teacher participants. Include the percentage of each group when available.
    1. 3-5 years :
    2. 5-10 years :
    3. Unknown: I**f the study gives averages and percentages for different ranges, please list those. Include the median if it is given.**
    4. 10-15 years
    5. 16 or more years

14. Highest degree held by teacher participants. Include the percentage of each group when available.
    1. Bachelor’s degree
    2. Master’s degree :
    3. Doctorate
    4. Other (explain)
    5. Unknown

15. Type of program being taught code all that apply
    1. General Ed
    2. Special Ed
    4. Migrant
    7. Other: _____________
    4. Title I
    5. GATE
    6. ELL

16. Code the subjects being assessed.
    1. Mathematics
    2. Reading
    3. Language Arts
    4. Writing
    5. Science
    6. Other

17. Describe the setting of the study. Include the geographic location of the study (state school or district placement: urban, suburban, rural, etc...) as well as type of school (private or public).

18. List the areas of Board Certification included in the study. (i.e., Early Childhood Generalist, Early Adolescent Mathematics, etc…)

Research Design Descriptors

19. Total student sample size. (Start of study if same throughout the study if not create a table with year, subject or grade level with table totals)

Year	Reading	Language Arts	Math
1999-2000			
2000-2001			
2001-2002			
2002-2003			

20. Student sample size by grade level and certification type. (Start of study)
    1. 3-5 grade:
    2. 6-8 grade

3. 9-12 grade
4. **Create a table for multiple years, grade levels, and tests**

Year	Grade Level	Reading	Language Arts	Math
1999-2000	3			
	4			
	5			
	6			
2000-2001	3			
	4			
	5			
	6			

21. Total teacher sample size. (Start of study)
Create a table for multiple years, grade levels, and tests

22. Teacher sample size by certification type. (Start of study)
1. National Board Certified Teacher sample size.
2. Nonboard Certified Teacher sample size. (both teachers who attempted board certification and either did not complete the process or did not become certified and teachers who have never attempted board certification.)

23. Was there a pre-test?
1. yes　　2. no

24. If there was a pre-test, were groups tested for equivalence using the pre-test? (look for ANOVA or testing on pretest only)
1. yes　　2. no

25. Type of assignment to condition.
1. matching　　4. blocking
2. random　　5. nonrandom
3. unknown

26. Code all research designs.
1. regression
2. independent samples t-test
3. Analysis of covariance (ANCOVA)
4. Analysis of variance (ANOVA)
5. Multivariate analysis of variance (MANOVA)
6. Hierarchical Linear Model (HLM)
7. Ordinary Least Squares (OLS)
8. Nested variables (list nested variables)
9. Value-added

27. Fixed effects (list all control variables used)
a. individual

        1. gender
        2. ethnicity
        3. free and reduced lunch
        4. ESL programming
    b. classroom
        1. class size
        2. % receiving free or reduced lunch
        3. % Caucasian
    c. School
        1. school size
        2. % receiving free or reduced lunch
        3. %Caucasian, Hispanic, Black

Nature of the Assessment Descriptors

28. Name of test used ____.

29. Code the type of assessment
    1. National exam
    2. State exam
    3. Local test

EFFECT SIZE LEVEL CODING MANUAL

For each effect size, code all of the following items

1. Study ID number. Identification number of the study from which the effect size is coded.

2. Round statistic data to two decimal places.

3. Effect size number[ESNUM]. Assign each effect size within a study a unique number. Number multiple effect sizes within a study sequentially, e.g. 1, 2, 3, 4, etc…
   Create a table for multiple years, grade levels, and tests

Dependent Measure Descriptors

4. Effect size type. [ESTTPE] Code an effect size as a gain score if the measures being compared across groups are the differences between the one year's scores and the previous year's scores. Code pre/posttests if the measures were taken in the same school year. Single comparison effect sizes are based on comparisons of single scores between groups. Code all scores taken over multiple years sequentially, e.g. 1, 2, 3, 4, etc…
   Create a table for multiple years, grade levels, and tests
    1. gain scores

2. pretest/posttest
3. end of course/grade scores
4. Outcome descriptor ________________________

5. Outcome descriptor.
   a. Norm referenced - **Most state and national tests considered norm referenced**
   b. Curriculum based –Bases on **curriculum used in the classroom.**
   c. Criterion referenced **– Based on standards but not subject to the normative process**

Effect Size Data

6. Type of effect size based on
   a. means and standard deviations
   b. t- value or F- value
   c. chi-square (df=1)
   d. frequencies or proportions, dichotomous
   e. frequencies or proportions, polychotomous
   f. gain scores
   g. other (specify) ___________________________

7. Page number(s) of where effect sizes were found ___________.

8. Raw difference favors (i.e., shows more success for which group?)

If the 2 groups are different in performance with different populations give the numbers
and percents for the overall of each group.
Create a table for multiple years, grade levels, and tests

0. Neither
1. National Board Certified Teachers
2. Nonboard Certified Teachers

Year	Grade	Raw difference favors			Data		
1999-2000	3rd	Rdg	LA	Math	Rdg	LA	Math
		Higher gain scores NBCTs	Higher gain scores NBCTs	Higher gain scores NBCTs	SD same for both groups	SD lower for NBCTs	SD lower for NBCTs
	4th	Same gains for both groups	Same gains for both groups	Higher gain scores for non-NBCTs	SD somewhat higher for NBCTs	SD lower for NBCTs	SD lower for NBCTs

When means and standard deviations are reported or can be estimated

9a. Student sample size (write in the appropriate number) and Group means

Create a table for multiple years, grade levels, and tests

1. National Board Certified Teachers
2. Nonboard Certified Teachers

9b. Group means

Create a table for multiple years, grade levels, and tests

1. National Board Certified Teachers
2. Nonboard Certified Teachers

9c. Group Standard Deviations

Create a table for multiple years, grade levels, and tests

1. National Board Certified Teachers
2. Nonboard Certified Teachers

When frequencies or proportions are **if this is the only thing reported or is all that can be estimated create a table for multiple years, grade levels, and tests**

10a. n of each group

1. National Board Certified Teachers
2. Nonboard Certified Teachers

10b. Proportion of group with successful outcomes (write in the value, if available)

Successful outcomes are indicated by higher frequencies or proportions of gain scores if the results are different for certain grades or certain subjects then create a table with the differences. When available include the overall proportion of successful outcomes.

1. National Board Certified Teachers
2. Nonboard Certified Teachers

STUDY LEVEL CODING PROTOCOL

Coder Initials:

Biographic Reference

Write a complete citation in APA form

1. Study ID Number [STUDYID]

2. Publication ID Number [PUBID]

3. Type of publication [PUBTYPE]
   1 Peer reviewed
   2 Dissertation
   3 Unpublished Conference paper
   4 Web based
   5 Other:

4. Publication year [PUBYEAR]

Sample Descriptors

5. Mean age of the student sample [MEANAGESTU]

6. Grade levels [GRADELEVELS]
   1. Elementary (k-5)
   2. Elementary (k-6)
   3. Middle School (6-8)
   4. High School(9-10)
   5. Write in grade levels if different from above list ____________________

7. Race of student sample [RACESTU]
   1. Caucasian:
   2. African Am.:
   3. Asian .:
   4. Hispanic:
   5. Native Am.:
   6. Unknown:

8. Gender of student sample. [GENDERSTU]
   1. Male:
   2. Female:
   3. unknown

9. Socio-economic status of student sample. [SESSTU]
   1. affluent
   2. middle class
   3. Poor.
   4. unknown.

10. Gender of teacher sample. [GENDERTEACH]
    1. Male:
    2. Female:
    3. unknown

11. Race of teacher sample. [RACETEACH]
1. Caucasian:
2. African Am.:
3. Asian .:
4. Hispanic:
5. Native Am.:
6. Unknown:

12. Teaching experience [TEACHEXP]
1. 3-5 years :
2. 5-10 years :
3. unknown
4. 10-15 years
5. 16 or more years

13. Degree [DEGREE].
1. Bachelor’s degree
2. Master’s degree
3. Doctorate
4. Other (explain)
5. Unknown

14. Program [PROGRAM]
1. General Ed
2. Special Ed
4. Migrant
4. Title I
5. GATE
6. Other: _____________

15. Subjects assessed [SUBJECT]
1. Mathematics
2. Reading
3. Language Arts
4. Writing
5. Science
6. Other

16. Setting of the study.
1. Urban
2. Suburban
3. Metropolitan
4. Rural
5. Private
6. Public
7. Other ______________________

17. List the areas of Board Certification included in the study. (i.e., Early Childhood Generalist, Early Adolescent Mathematics, etc…)
____
___

Research Design Descriptors

18. Total student sample size. [TOTALN]

19. Student sample size by grade level and certification type. [GRADELEVELN]
1. 3-5 grade:
2. 6-8 grade
3. 9-12 grade

20. Total teacher sample size. [TEACHERN]

21. Certification type. [CERTIFICATION]
1. National Board Certified Teacher sample size.

2. Nonboard Certified Teacher sample size.
3. Never attempted Board Certification Teacher sample size

22. Was there a pre test?
    1. yes 2. no

23. If there was a pretest, were groups tested for equivalence using the pretest? [PREEQUIV]
    1. yes 2. no

24. Type of assignment to condition. [ASSIGN]
    1. matching
    2. random
    3. unknown
    4. blocking
    5. nonrandom

25. Research design. [DESIGN]
    a. regression
    b. independent samples t-test
    c. Analysis of covariance (ANCOVA)
    d. Analysis of variance (ANOVA)
    e. Multivariate analysis of variance (MANOVA)
    f. Hierarchical Linear Model (HLM)
    g. Ordinary Least Squares (OLS)
    h. Nested variables (list nested variables)
    i. Value-added

26. Fixed Effects
    1. levels (list for each area)
    2. individual
    3. classroom
    4. school

Nature of the Assessment Descriptors

27. Name of test used ____.

28. Code the type of assessment [ASSESSMENT]
    1. National exam
    2. State exam
    3. Local test

EFFECT SIZE LEVEL CODING PROTOCOL

For each effect size, code all of the following items

1. Study ID number. [STUDYID]

2. Effect size number. [ESNUM]

Dependent Measure Descriptors

3. Effect size type. [ESTYPE]
   1. gain scores
   2. pretest/posttest
   3. end of course/grade scores

4. Outcome descriptor ____

Effect Size Data

5. Type of effect size based on [ESDATA]
   1. means and standard deviations
   2. t- value or F- value
   3. chi-square (df=1)
   4. frequencies or proportions, dichotomous
   5. frequencies or proportions, polychotomous
   6. gain scores
   7. other (specify) __________________________

6. Page number(s) of where effect sizes were found [PAGENUM] ___________.

   Use the article page number and not the pdf page numbers.

7. Raw difference favors (i.e., shows more success for [SUCCESS]

   If the 2 groups perform differently
   1. National Board Certified Teachers
   2. Nonboard Certified Teachers

When means and standard deviations are reported or can be estimated

7a. Student sample size [STUN] and Group means [GRPMEAN]
   1. National Board Certified Teachers
   2. Nonboard Certified Teachers

7b. Group Standard Deviations [GRPSD]
   1. National Board Certified Teachers
   2. Nonboard Certified Teachers

When frequencies or proportions are reported or can be estimated

8a. n of each group
   1. National Board Certified Teachers [NBPTN]
   2. Nonboard Certified Teachers [NNBPTSN]

8b. Proportion of group with successful outcomes (write in the value, if available)

3. National Board Certified Teachers [NBPTSUCCES]
4. Nonboard Certified Teachers [NNBPTSSUCCES]

General Directions:

- Include page numbers on coding sheets. Use the article page number and not the pdf page numbers.
- Non-board certified includes both teachers who went through the process but did not certify and those who never attempted certification. For studies that separate them into 2 distinct categories collect the data separately in a table to be combined later for statistical analysis
- Do not combine data from separate questions in tables. Create tables that address each question.
- Example of a table
  1999-2000

	Grade 3	Grade 4	Grade 5
Reading	**1** non-NBCTs **1.21** **2** NBCTs **1.35**	**7** non-NBCTs **1.78** **8** NBCTs **1.28**	**13** non-NBCTs **0.55** **14** NBCTs **0.83**
Math	**3** non-NBCTs **1.00** **4** NBCTs **1.41**	**9** non-NBCTs **1.25** **10** NBCTs **1.23**	**15** non-NBCTs **0.99** **16** NBCTs **1.73**
Language	**5** non-NBCTs **1.04** **6** NBCTs **1.20**	**11** non-NBCTs **0.84** **12** NBCTs **0.93**	**17** non-NBCTs **0.49** **18** NBCTs **0.71**