12-31-2013

The Association Between Elementary Teacher Licensure Test Scores and Student Growth in Mathematics: An Analysis of Massachusetts MTEL and MCAS Tests

Life LeGeros
University of Massachusetts Boston

Recommended Citation

THE ASSOCIATION BETWEEN ELEMENTARY TEACHER LICENSURE TEST SCORES

AND STUDENT GROWTH IN MATHEMATICS:

AN ANALYSIS OF MASSACHUSETTS MTEL AND MCAS TESTS

A Dissertation Presented

by

LIFE LEGEROS

Submitted to the Office of Graduate Studies,
University of Massachusetts Boston,
in partial fulfillment of the requirements for the degree of

DOCTOR OF EDUCATION

December 2013

Leadership in Urban Schools Program

THE ASSOCIATION BETWEEN ELEMENTARY TEACHER LICENSURE TEST SCORES

AND STUDENT GROWTH IN MATHEMATICS:

AN ANALYSIS OF MASSACHUSETTS MTEL AND MCAS TESTS

A Dissertation Presented

by

LIFE LEGEROS

Approved as to style and content by:

_____
Wenfan Yan, Professor
Chairperson of Committee

_____
Michael Gilbert, Assistant Professor
Member

_____
Erin O'Brien, Associate Professor
Member

_____
Tricia Kress, Program Director
Leadership in Urban Schools Program

_____
Wenfan Yan, Chairperson
Department of Leadership in Education

ABSTRACT

THE ASSOCIATION BETWEEN ELEMENTARY TEACHER LICENSURE TEST SCORES

AND STUDENT GROWTH IN MATHEMATICS:

AN ANALYSIS OF MASSACHUSETTS MTEL AND MCAS TESTS

December 2013

Life LeGeros, B.A., Grinnell College
M.A., University of Madison Wisconsin
Ed.D., University of Massachusetts Boston

Directed by Professor Wenfan Yan

This quasi-experimental value-added study provided evidence for the predictive

validity of the Massachusetts MTEL General Curriculum Mathematics Subtest by finding

an association between the licensure test results of 130 teachers and the growth of their

2640 grade 4 and 5 students. The study took advantage of a natural experiment that arose

due to a policy change made by the Massachusetts Department of Elementary and

Secondary Education (MADESE) in response to the initial administration of a new highly

rigorous math-specific licensure subtest for elementary and special education teachers in

March, 2008. The emergency amendment allowed test takers to conditionally pass the

licensure test based upon a lower, temporary cut score, therefore providing a comparison

group of teachers who received conditional licensure without fully passing the licensure

test. The study sample used a cross-sectional data set acquired from MADESE for the

2010-11 school year, the first year for which data was available that linked individual

teachers to their students. The dependent variable of students' mathematics Student

Growth Percentile (SGP) score on the statewide test, the MCAS, incorporated prior

achievement and was calculated by comparing each student to his or her academic peers.

OLS regression analyses including student background variables, classroom variables, and

teacher characteristic variables showed that teacher results on the MTEL math test were

positively associated with student math SGP scores. The strength of the association found

in this study was substantial relative to the research literature and comparable in

magnitude with established factors such as student low-income status. The predictive

power of the MTEL math test was strongest at the lower range of test scores, suggesting

that policymakers should consider lowering the permanent cut score to the level set by the

emergency amendment in order to avoid screening effective teachers out of the workforce

and potentially decreasing student achievement.

ACKNOWLEDGMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Every student deserves a good teacher, even if that teacher is brand new to the

job. Policymakers are obligated to ensure that new teachers are at least minimally

competent when they enter the profession. States use licensure as a policy mechanism to

meet this obligation: "the specified purpose of teacher licensure is to guarantee a

minimum standard of quality of public school teachers by avoiding the possibility that

poor local hiring decisions result in the employment of unsuitable teachers" (Goldhaber,

2004, p. 83). To protect students and the public good, then, states must define and

measure a minimal standard of teacher quality. If the definition or measure is flawed,

however, licensure policies could potentially have adverse effects on student

achievement.

The Massachusetts Department of Elementary and Secondary Education

(MADESE) requires teacher candidates to possess a Bachelor's degree and pass a series

of paper-and-pencil tests to earn the entry-level ("preliminary") license. The state has

recently changed the teacher tests for elementary teachers (including K-8 special

education teachers) to increase the requirements in mathematics by including a newly

developed mathematics subtest. As a gatekeeper into the profession, the mathematics

subtest represents a particular definition and measure of teacher quality. Boyd et al

(2007) succinctly describe what is at stake in teacher licensure exams:

If the exams identify good teachers more effectively than hiring authorities can in the absence of the exams, then they could improve student achievement. But if the exams make distinctions based on knowledge that is not closely related to student outcomes, or if they classify individuals erroneously, they could exclude applicants who would be more effective teachers, thereby reducing student outcomes. (p. 54)

The question at hand, therefore, is whether the mathematics subtest defines and measures teacher quality in a way that is likely to improve student achievement, which is MADESE's primary concern.

This study examines the link between the mathematics subtest and student outcomes. Specifically, it evaluates the predictive validity of the mathematics subtest by investigating the association between teacher outcomes on the mathematics subtest and student outcomes on the state's high-stakes mathematics assessment.

Problem

Although Massachusetts students perform well overall in national and international comparisons, persistent achievement gaps and a growing number of schools identified for improvement in mathematics are indicative of systemic challenges, as described by the MAESE Commissioner in a presentation titled *Celebrating Progress, Committing to Next Steps for Narrowing Achievement Gaps* (Chester, 2008). On the 2007 National Assessment of Education Progress, for example, Massachusetts' fourth graders were first in the country in mathematics but ranked fourth for African-American students and fifteenth for Hispanic students. In 2012, the state's test, the MCAS, reflected those same gaps with a 25-point difference for 10th graders between White

students and African-American students and a 29-point gap between White students and Hispanic students, compared to 20-point gaps for each group in 2000, and the gaps between regular education students and students with disabilities as well as Limited English Proficient (LEP) students and non-LEP students widening during that same period. Most strikingly, the number of schools in federal accountability status for has risen steadily since the passage of NCLB, with more than half of all schools identified, mostly due to mathematics performance, and a staggering 95% of urban middle schools identified for improvement. The overall message of Chester's presentation was that although much has been accomplished there is still much work to do in Massachusetts.

MADESE has a wide array of policies and initiatives to address the achievement gap and school improvement challenges, many of which focus on improving the current workforce. The licensure test policy is directed at improving the pipeline of incoming teachers. The licensure tests have high-stakes consequences for teachers and therefore demand a high standard of evidence for justification. Unfortunately, the existing research base on licensure policies is generally inconclusive, with some studies linking teacher certification to student achievement (Clotfelter, Ladd, & Vigdor, 2007, 2010; Neild, Farley-Ripple, & Byrnes, 2009) and others finding no association (Betts, Zau, & Rice, 2003; Croninger, Rice, Rathbun, & Nishio, 2007; Goldhaber & Brewer, 2000; Jepsen, 2005; Phillips, 2010).

Research on teacher testing is similarly mixed. Although some studies have found teachers' performance on licensure tests to be associated with their students' achievement (Clotfelter, Ladd, & Vigdor, 2006; Clotfelter, et al., 2010; Ferguson, 1991; Goldhaber & J. Hannaway, 2009), in most cases the teacher tests are less predictive than

other factors such as experience. One recent meta-analysis, for example, found that preservice preparation program performance (measured by grade point average) is a significantly better predictor of teacher competence than teacher licensure test scores (D'Agostino & Powers, 2009). Other studies have found no link between teacher tests and student achievement (Buddin & Zamarro, 2009) or have shown the association to vary widely depending on grade level and subject (Clotfelter, et al., 2007). The literature also suggests that teacher tests produce inequitable effects of various types, including possible test biases against minority teacher candidates (Bennett, McWhorter, & Kuykendall, 2006; Skiba, et al., 2008), uneven predictive validity based on the race of teachers (Goldhaber & Hansen, 2010), and decreased diversity of the teacher workforce (Angrist & Guryan, 2008). Although teacher tests are ubiquitous with 46 out of 50 states requiring teacher candidates to pass tests to receive licensure (Rotherham & Mead, 2004), the increase in rigor represented by the mathematics subtest is a bold policy move that is not necessarily supported by the research literature. The policy is problematic to the extent that it has unintended adverse effects or fails to provide information about suitability to teach beyond what is already available.

What we do know is that the substantial variations in student growth between classrooms have led to a consensus in the research community that teachers significantly impact student achievement (Hanushek & Rivkin, 2010a). The MADESE policy stands on somewhat shaky research ground because the research community is less sure about *how* teachers impact student achievement, and therefore how to predict which teachers will be competent. As is often the case, the stakes are even higher for urban areas due to the increased sensitivity to teacher quality of students with diverse backgrounds

(Aaronson, Barrow, & Sander, 2007; Phillips, 2010), as well as the likelihood that decreasing the pool of licensed teachers would more negatively impact urban districts (Podgursky, 2005). It is important that the MADESE mathematics subtest is measuring teacher quality meaningfully, and evaluation of the predictive validity of the test is necessary in light of the lack of a solid research basis for the policy.

Context

The policies of standards-based education reform primarily focus on measurable outcomes of the education process (Fuhrman, 2004). Policymakers define expectations, most centrally in the form of learning standards for (all) students, and then hold educators accountable for meeting the expectations. Failure to meet these standards can trigger high-stakes consequences for students (e.g., denial of high school graduation), schools (e.g., sanctions and public shaming), and teachers (e.g., termination). The theory of action is that the motivation to avoid these negative consequences will spur educators to improve practice and leaders to align systemic components such as student expectations and teacher preparation, leading to better outcomes for students in terms of both excellence and equity (O'Day & Smith, 1993). The emphasis on outcomes requires the quantification of success, which can pose a problem because "many valued educational objectives cannot be captured for measurement" (Moller, 2009, p. 40). Critics charge that emphasizing measurable outcomes narrows the purposes of education to its most quantifiable aspects (Schoen & Fusarelli, 2008) and ignores supporting inputs to force front-line practitioners to unfairly bear the full burden of responsibility for success (Cochran-Smith, 2005). Regardless of critiques along these and other lines, however, standards-based reform has become thoroughly entrenched in the K-12 arena and teacher

5

testing can be construed as the extension of this approach to regulate the outcomes of teacher preparation institutions (Tellez, 2003; Zuzovsky & Libman, 2006).

Simultaneously, teacher licensure test policies allow the possibility that teacher candidates can bypass teacher preparation institutions altogether. Zeichner (2003) dubbed the agenda of standards-based reformers "deregulation" because it forms the backbone of alternative routes to licensure. Although extreme deregulators argue for the abolition of licensure altogether (for example, see Podgursky, 2005), most proponents acknowledge the need to protect the public good by screening out obviously incompetent prospective teachers (Hess, Rotherham, & Walsh, 2004). Deregulators emphasize subject matter expertise as central to quality teaching and, even more importantly, objectively measurable in a way that lends itself to state oversight (Walsh, 2004). Although the deregulation agenda does not logically entail an outright rejection of the importance of unmeasurable skills and dispositions, in practice it devalues the substance of traditional teacher education programs. Critics of deregulation charge that it is an oversimplified view, based on outdated characterizations of teacher education, blind support for alternative routes to teaching, and obsession with subject matter knowledge at the expense of other forms of teacher knowledge such as pedagogy and cultural understanding (Zeichner, 2003).

In contrast to deregulation, the professionalization agenda advocates the establishment of teaching as a profession through the articulation of a consensus view of quality teaching based on education research and professional judgment (for example, see Darling-Hammond, Bransford, Lepage, Hammerness, & Duffy, 2005). Professionalization is not a defense of the status quo, and its advocates have fully

recognized the teacher quality problem and long called for substantial structural changes to teacher preparation and policy (for example, see Darling-Hammond, 1997). Teacher preparation suffers from the same inconclusive research basis as teacher licensure and testing that was noted above, and deregulators have been quick to point out that professionalization is based on weak evidence at best that may be presented in misleading ways at worst (for example, see Ballou & Podgursky, 2000). Other critics have charged that the yearning for consensus and legitimacy has caused professionalization to give short shrift to social justice issues (Zeichner, 2003), though more recently the teacher education establishment seems to have fully embraced multiculturalism and equity (for example, see Sleeter, 2008; Villegas & Davis, 2008). Deregulators tend to lump together professionalization and multiculturalism as touchy-feely approaches that are on their way out of vogue as pragmatism takes a firm hold due to the hard work of reformers (for example, see Farkas & Duffett, 2010).

Although some observers claim that this debate has recently reached a new level of vitriol (Hess, 2005), self-dubbed reformers have been at work in Massachusetts for some time and a sketch of teacher testing in Massachusetts reveals particulars about the overarching controversy. The Massachusetts Education Reform Act of 1993 was a model standards-based reform omnibus bill that included requirements for entry-level teacher testing, along with other systemic elements such as high-stakes testing and school-base accountability. Soon after the MADESE began testing in 1998, a series of commentaries by teacher educators catalogued a wide range of issues with the way that the tests were implemented and covered by the media (Flippo & Riccards, 2000; Fowler, 2001; Harrington, 1999; Melnick & Pullin, 2000). Specifically, the original Massachusetts

teacher tests have been criticized due to the teacher bashing associated with implementation, the poor quality of the test themselves, and negative impacts on teacher preparation institutions.

The first Massachusetts Educator Certification Test (MECT), administered in April, 1998, resulted in a reported 40% pass rate that spawned a media furor about the low knowledge base of prospective teachers. Two weeks beforehand, MADESE announced that the test would determine whether candidates received licenses, a change of course from the original plan that the first two testing sessions would be pilots without consequences for licensure. Although candidates were only counted as passing if they passed three distinct tests (reading, writing, and subject matter), and the cumulative pass rate after two years was over 70% (Fowler, 2001), some standards-based reformers pointed to the pass rate as evidence of the dire state of teacher preparation. The local and national media ran with the storyline that teachers who failed the test were "idiots", as declared by Speaker of the House Tom Finneran, and that the test was at the 10th grade level, so described by education Chairman John Silber in a New York Times oped article (as cited in Fowler, 2001). This characterization of the test seems misleading when it is taken into account that the MADESE's contract with the test developer, National Education Service, stipulated that the test should be at the college level and that an independent evaluation lauded the Massachusetts tests for going beyond other similar tests that were pegged at the high school level (Ruth & Barth, 1999). The MECT experience perhaps illustrates the oversimplification in which deregulators sometimes indulge.

The problems with the MECT were compounded by evidence that suggested that it was not a high quality assessment. There was abundant anecdotal evidence along these lines, such as a well-published Ph.D. holder from MIT who failed with a score of 59 on the initial administration and then passed with a 93 on her re-take (Fowler, 2001). There was also highly questionable content such as an infamous dictation task that counted for 25% of the writing test in which candidates were scored on spelling, punctuation, and capitalization as they scribed an 18th century passage played from an audio tape (Melnick & Pullin, 2000). The Center for the Study of Testing, Evaluation and Educational Policy at Boston College formed an ad hoc committee that evaluated MECT validity and reliability using professional standards for testing (Haney, Fowler, Wheelock, Bebell, & Malec, 1999). The committee reported that the margin of error was two to three times larger than was acceptable, resulting in high false-pass and false-failure rates. The tests did not correlate well with established measures and were highly unreliable, with candidates' scores swinging widely at different sittings. The panel recommended that the state suspend the testing program and convene an independent panel to audit the tests, but an MADESE promised independent review never materialized and the testing vendor offered technical reports that were not adequate or convincing to critics (Fowler, 2001).

If indeed the MECT were bad tests, they would negatively impact teacher education to the extent that preparation institutions rearranged their programs to "teach to the test." Flippo (2000) argued that the MCET would decrease the teacher pool both generally because of the degrading way that teachers were treated, and specifically for nonwhite prospective teachers. MCET also may not have incentivized the type of change in teacher preparation that was intended. For example, it is common to this day for

teacher preparation institutions to game the system by simply requiring candidates to pass the tests before they can complete the practicum portion of their programs, thus ensuring a 100% pass rate for their graduates. The most alarming negative impact, however, is the possibility that issues with assessment quality caused teachers to be screened out of the profession in error. Fowler (2001) noted the blind faith in the tests that was exhibited by the public, policymakers, and the media. Although paper-and-pencil tests are attractive to reformers in large part because of their objectivity, this advantage is meaningless if the measurement instrument is invalid.

The initial implementation of teacher testing in Massachusetts is representative of many of the issues surrounding teacher testing specifically as well as the conflict between the dominant standards-based reform agenda driven by deregulators and the professionalization agenda championed by teacher educators. Although the account here is assembled largely from the viewpoint of teacher educators that likely support professionalization, it is apparent that regardless of any missteps on the part of reformers in the way of teacher bashing, reliance on an invalid assessment, or unintended consequences of the policy, the reformers are firmly in control of the state education machinery. Massachusetts currently only requires a Bachelor's degree in addition to passing the teacher tests in order to gain an entry-level license. Therefore the same concerns from the professionalization side still hold today and it is incumbent on the state to produce evidence that the teacher tests, as sole gatekeeper, are valid.

As the only state to require a mathematics subtest for elementary teachers (Greenberg & Walsh, 2008; Stotsky, 2009), Massachusetts once again finds itself at the leading edge of standards-based reform. In April 2007 the Board of Elementary and

Secondary Education strengthened the math content requirements for elementary and special education teachers and in March 2008 administered the mathematics subtest for the first time (the state's licensure tests continue to be developed by National Education Service). The initial pass rate of 27% at the scaled cut-score of 240 forced the Board to enact an emergency amendment that allowed candidates with a scaled score of 227-239 to receive preliminary licensure until June 2012. The new cut score increased the pass rate to approximately 40%, and subsequent administrations have been closer to 50%, compared to rates near 70% on the previously combined subject matter test. The Board must carefully consider the impact of the new teacher testing policy on its standards-based reform agenda, which ultimately seeks to increase overall student achievement and close achievement gaps (Chester, 2009a). This study will look at the predictive validity of the mathematics subtest, thus speaking to policy makers directly in the language of measurable student outcomes that they prefer, and contributing to the evidence base that is referenced by members of both the deregulation and professionalization agendas.

Rationale

One recent review of the literature on teacher preparation and certification ultimately concluded that "the research evidence is simply too thin to have serious implications for policy" (Boyd, Goldhaber, Lankford, & Wyckoff, 2007). Therefore, a primary audience for this study is policymakers. As noted above and put eloquently by Imig (Imig & Imig, 2006), measurability is key to standards-based policies such that "subject matter knowledge and student achievement are the currency of the realm in which we must operate" (p. 168). This study essentially uses standards-based reform's logic to evaluate one of its policies by linking subject matter knowledge measured by

11

teacher tests to student achievement measured by the high-stakes state assessment. Massachusetts is particularly interesting to policy makers due to its status as a relatively high-performing state with strong reform credentials.

Urban policy makers and educators are an especially relevant audience for this study. Teacher quality in general is a particularly salient issue in urban school systems that often struggle to hire high quality teachers (Darling-Hammond, 2007; Lankford, Loeb, & Wyckoff, 2002; Podgursky, 2005; Stotko, Ingram, & Beaty-O'Ferrall, 2007) or distribute teachers in a way that doesn't disadvantage nonwhite and poor students (Boyd, Lankford, Loeb, Rockoff, & Wyckoff, 2008; Clotfelter, et al., 2010). Students with diverse backgrounds tend to be more sensitive to teacher quality and the impacts of effective teachers are powerful enough to potentially offset at-risk factors (Heck, 2007; Phillips, 2010; Sanders & Horn, 1998). There is a body of literature that suggests that teaching in urban schools requires a broad and unique skill set that is extremely difficult to capture in a paper-and-pencil format like the math subtest (Delpit, 2006; Ladson-Billings, 1998; Stotko, et al., 2007). A recent study of North Carolina teacher tests possibly supports this view with evidence that, on average, black teachers who score below the cut-score on the states licensure exams teach black teachers more effectively than white teachers who score above the cut-score (D. Goldhaber & Hansen, 2009). If potentially effective teachers of diverse students were denied licenses, urban schools and districts would bear the brunt of the negative consequences.

The research community is an important audience for this study as well, with researchers on both the professionalization and deregulation sides of the debate interested in quantitative evidence related to teacher testing. This study offers a unique opportunity

to contribute to the literature for a number of reasons. First, the focus on elementary

mathematics addresses a hot topic in policy circles. A recent survey study of teacher

preparation institutions by the National Center on Teacher Quality finds that only 10 of

77 elementary teacher preparation programs include adequate mathematics content

preparation (Greenberg & Walsh, 2008). The authors advocated stand-alone mathematics

licensure tests for elementary teachers and noted (with much anticipation) that

Massachusetts is the first state to put in place such a test. Second, elementary

mathematics is an interesting intersection from a research standpoint, since the literature

tends to show a stronger influence of teacher content knowledge on student achievement

in mathematics than other subjects, but mostly at the high school level (Wayne &

Youngs, 2003). There is a solid basis supporting pedagogical content knowledge at the

elementary level (Hill, Rowan, & Ball, 2005), however, and the Massachusetts

mathematics subtest purposefully includes items modeled on these studies. It is an open

question whether a rigorous measure of mathematics knowledge at the elementary level,

slightly more broadly defined than pure subject matter knowledge, will show connections

to student achievement.

The third unique aspect of this study derives from the quasi-experiment set up by

the emergency amendment that effectively creates natural comparison groups of (a)

teachers who received "emergency" licensure by reaching the conditional cut score and

(b) teachers who fully passed by reaching the permanent cut-score. It is usually difficult

to study the effectiveness of teachers who fail licensure tests because they have a hard

time getting teaching jobs, especially at the elementary level. The policy has created a

four year period in which two different cut scores will be in place, the conditional passing

score and the permanent passing score, to temporarily create three different comparison groups and also a larger overall pool of teachers who are in classrooms without achieving the permanent cut score.

Finally, the data set is exceptional due to recent advancements in the MADESE's database that allows teachers to be linked to individual students. The 2011 administration of the high-stakes state test was the first time that teacher-student linked data was available at the state level and was coincidentally the first year that a full cohort of teachers licensed under the new mathematics subtest were in classrooms. Also, in 2009 Massachusetts put into place an innovative student growth measure based on methodology that compares students to the predicted growth of their academic peers. As of 2012, this "Student Growth Percentile" (SGP) metric has been incorporated into accountability determinations at the school and district levels and is a mandatory part of the new educator evaluation systems that must be in place for the 2013-14 school year (Patrick, Chester, & Banta, 2010). The state's use of SGP for these high-stakes pruposes suggests that it could provide a distinct perspective on student growth that augments the value-added modeling approach commonly used in teacher effectiveness studies (for example, see Hanushek & Rivkin, 2010a).

Conceptual Framework

Goe (2007) presented a conceptual framework for teacher quality that clarifies the focus of this study on easily measurable "paper characteristics" of teachers and their link to student achievement. This is justified due to the fact that the association between these inputs and student achievement is strongly present in the literature and because these teacher characteristics are readily available to the state through its current data

collection practices (and therefore represent an alternative to using licensure testing to gather more information).

Figure 1 shows the conceptual framework of Goe (2007). It included three categories, with each category including one or more ways of looking at teacher quality:

- o Inputs - Teacher qualifications and teacher characteristics
- o Processes - Teacher practices
- o Outcomes - Teacher effectiveness

The first two categories, which relate to "who" is teaching and "how" they teach, relate to teaching in a task sense, while the third category defines teaching in an achievement sense (see Fenstermacher & Richardson, 2005). Goe asked the reader to:

Note that teacher qualifications, characteristics, and practices are all used to define teacher quality and exist *independently* of student achievement, whereas teacher effectiveness is wholly dependent on student achievement. In other words, teacher effectiveness cannot be determined without outcomes such as standardized test scores. The other three ways of looking at teacher quality can be theoretically connected to student learning and measured with standardized test scores, but they exist whether or not they are measured. For example, teacher certification exists as a proxy for teacher quality, even if it is never connected to student outcomes. But teacher effectiveness exists only as a function of the link between teachers and their students' standardized test scores. (p. 9)

This way of framing teacher effectiveness resonates with the common sense (and achievement sense) notion at the heart of standards-based reform that a teacher hasn't

taught something until their students have learned it. A teacher is not effective unless their students have demonstrably learned the material that was taught.

Teacher qualifications are largely limited to things that can be ascertained on paper, such as coursework, grades, degrees, experience, licensure status, credentials, and, of course, teacher licensure scores. This teacher quality framework is particularly relevant to this study because of its connection to the mathematics subtest policy which is designed to screen teachers entering the profession. Paper qualifications are the most obvious way to screen teachers at the state level and this study will include all of the paper qualifications available in the state data: licensure tests results, licensure status, years of experience, undergraduate major, and graduate degree(s).

Teacher characteristics include attributes and attitudes along with immutable aspects such as race and gender. This study will not examine these immutable aspects because they are not directly addressed by the policy and the teacher sample is not large or varied enough to allow examination of test bias. Similarly, teacher practices are beyond the scope of this study because from the current state perspective, practices are a black box that cannot be objectively measured for the purpose of licensure. Although performance assessments for licensure are in place in many states, their apparent promise for broadening teacher testing has been hampered by their high cost (Youngs, Odden, & Porter, 2003).

The outcomes category includes teacher effectiveness, which Goe distinguishes from student achievement by the fact that it takes into account prior achievement to provide a measure of the value added by the teacher. This study will use student outcomes on the Massachusetts Comprehensive Assessment System high-stakes

16

Figure 1. Teacher Effectiveness Conceptual Framework

Teacher Qualifications

Processes

Inputs

Teacher Practices (Teaching Quality)

Teacher Characteristics

Teacher Quality

Student Achievement Test Scores
(treated as indicator
of teacher quality)

Outcomes

Student Achievement (predicted) – Student Achievement (actual) = Student Gain Score

Teacher Effectiveness
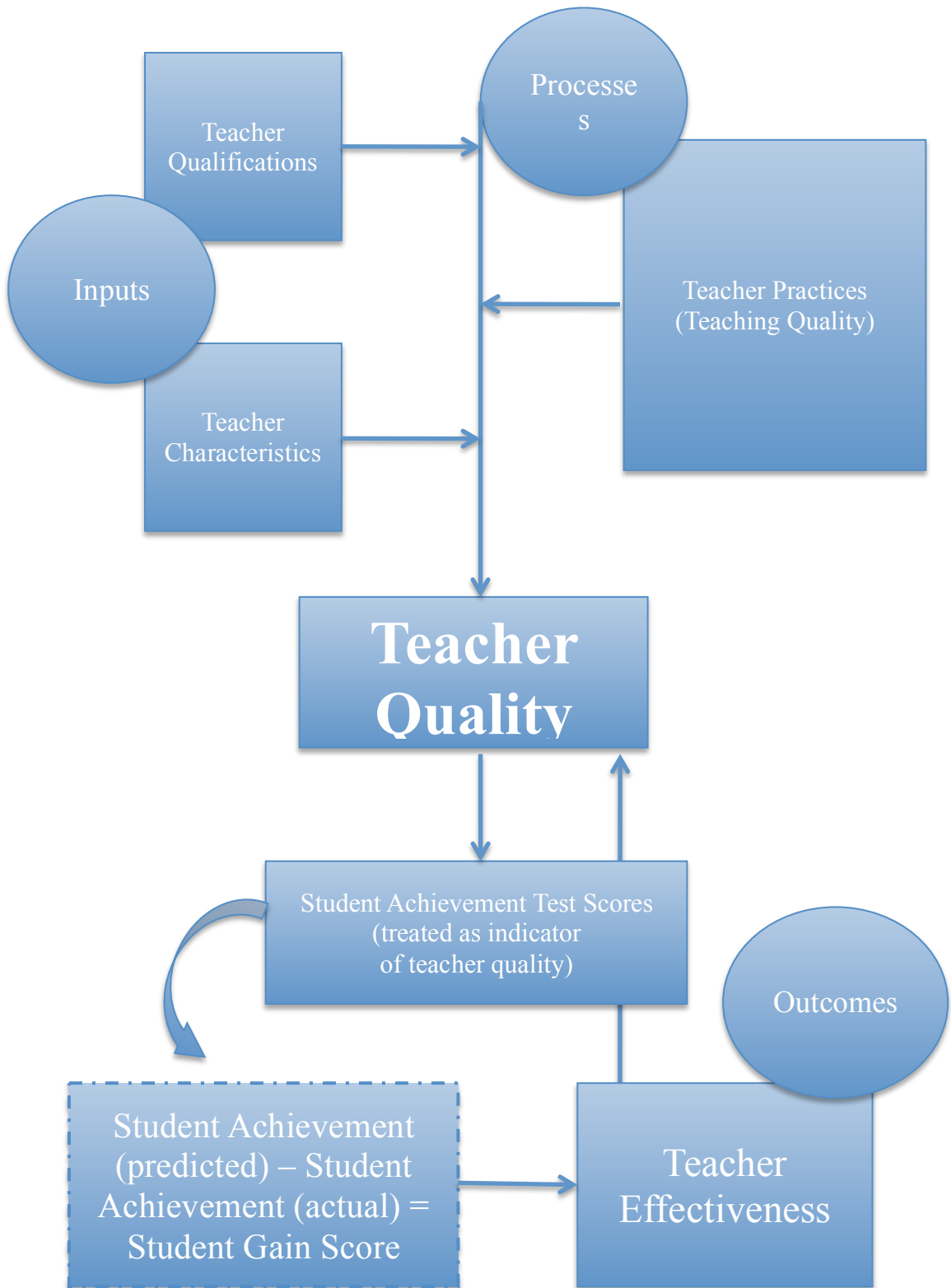
assessment for this purpose, along with the Student Growth Percentile statistic.

Goe is on song with standards-based reform in her advocacy of linking inputs and outcomes. Teacher effectiveness brings student achievement into the teacher quality model. This study acknowledges these various ways of understanding teacher quality, but limits the focus to those that are practically useful for screening teachers into the profession (i.e., paper qualifications) by using the data available to the state to make the study relevant to the policy dilemma at hand.

Research Questions and Methodology

This study generally employs a teacher effectiveness approach to examine the overarching research question, "Is the MA MTEL math subtest a valid predictor of teacher effectiveness?", along with these related research subquestions:

- Are teacher outcomes on the MTEL math subtest associated with student growth on MCAS, the high-stakes state test?

- How does the strength of the association between teacher outcomes and student outcomes, if any, compare with the strength of the association between other teacher characteristics (license status, graduate degrees, experience) and student growth?

- How does the strength of the association between teacher outcomes and student outcomes, if any, compare with the strength of the association between student characteristics (e.g., race, gender, low income status, English Language Learner status, special education status) and student growth?

Specifically, this study uses a cross-sectional cohort analysis of the group of 130 teachers who took the new MTEL mathematics subtest between March 2008 and July

2010, taught as full classroom teachers (i.e., not support or co-teachers) during the 2010-11 school year, and were linked to students. The study is cross-sectional because the study subjects are teachers and I use a data set that is effectively a snapshot from the 2010-11 school year. The main outcome measure is longitudinal, however, because SGP is calculated based upon past achievement. Prior achievement is also included as controls in some models for individual students or as peer effects.

The general teacher effectiveness framework is further explored in Chapter 2 and the methodology for this study is detailed in Chapter 3.

CHAPTER 2

LITERATURE REVIEW

The teacher effectiveness literature has surged in the last decade as researchers have sought to examine impacts on student achievement. This review first looks at teacher effectiveness in general by appraising the strong evidence that teacher quality impacts student achievement. Though it is well established that teachers measurably vary in terms of effectiveness, it is less clear what makes some teachers more effective than others. The second section of this review surveys the extensive literature on the impact of specific teacher inputs on student achievement. The final section of this review further considers historical context and issues specifically related to the new mathematics subtest and elementary teacher licensure policies in Massachusetts.

Teacher Effectiveness: The Impact of Teacher Quality on Student Achievement

As described in the introduction, a teacher's effectiveness can be defined by the impact that they have on student growth. Teacher quality is a broader, more multifaceted concept that encompasses processes as well as indicators such as qualifications and characteristics. Teacher effectiveness focuses on the student and tries to quantify the extent to which variation in student growth (as measured by a standardized test) can be attributed to teachers.

In a survey of results and methodological issues in teacher effectiveness research, Hanushek and Rivkin (2010a) summarized results from ten recent studies conducted

since 2004. The authors expressed the findings of each study in terms of units of student achievement and reported that within-school variation in the value added by teachers in mathematics ranged from 0.11 to 0.36 standard deviations of student achievement. They explained that this magnitude of impact implies that having a teacher at the twenty-fifth percentile as compared to the seventy-fifth percentile of the quality distribution would make the difference of approximately 0.2 standard deviations in a single year, moving a student from the middle of the achievement distribution to the fifty-eighth percentile. By comparison, the authors noted that convincing estimates of ten student class size reductions show effects of 0.1-0.3 standard deviations. The authors concluded that although teacher effectiveness research has room for methodological improvement, the consistency of practically significant effect sizes across so many studies that all employ sophisticated statistical techniques with strong data sets provides the basis for the airtight assertion that "teacher quality is an important determinant of school quality and achievement" (p. 268).

Value Added Modeling

All of the teacher effectiveness studies analyzed in Hanushek and Rivkin (2010) utilized value added modeling statistical techniques. It is instructive to examine how teacher effectiveness research is situated in the broader field of teacher education research and then consider the basic features of education production functions, the use of control variables to isolate the value added by teachers, and how researchers have grappled with the challenge of accounting for the nonrandom sorting of students into classrooms (and therefore, systematic matching of students and teachers).

Teacher education research genres and paradigms. Teacher education is successful to the extent that the teachers that it produces are effective. This commonsense view has allowed teacher effectiveness research to become a dominant force in teacher education research. In the most recent Handbook of Research on Teacher Education, David Imig, the former President and CEO of the American Association of Colleges of Teacher Education declared "education professionals have to embrace an agenda of teacher effectiveness . . . or their voice will be lost in the debate" (Imig & Imig, 2008, p. 904). In the Introduction, I considered the ideological differences between the professionalization advocates and the deregulators who emphasize teacher effectiveness. The final section of this literature review looks at the historical context of the shift to deregulation, and corresponding emphasis on teacher licensure testing. These policy shifts are mirrored by the ascension of a teacher education research genre that emphasizes effects and a policy-focused research paradigm that uses methods from economics.

Borko, Whitcomb, and Byrnes (2008) argued that two genres of teacher education research are well established and dominant: the interpretive and effects genres. Interpretive research commonly employs qualitative methods and is the "most expansive category . . . [including] ethnography, symbolic interactionism, narrative, educational connisseurship, phenomenology, and discourse analysis" (p. 1024). The effects genre tends to use quantitative methods from the field of economics to look at the factors that impact student outcomes. Though experimental studies are the gold standard of this genre, they are rare and researchers more commonly employ quasi-experimental approaches involving the analysis of large data sets with sophisticated correlational

statistical techniques. Critics charge that the emphasis on objective and generalizable results is misguided in the field of education due to the intrinsically personal and contextual nature of teaching and learning (Florio-Ruane, 2002). As the preferred research genre of policymakers, however, quantitative approaches have been largely rebranded as "scientifically-based research" and teacher effects research has received funding and increased status by the standards-based reform movement (Borko, et al., 2008).

Using a research paradigm framework, Cochran-Smith and Fries (2008) examined 38 syntheses of research literature and traced the way that "the problem" of teacher education has been defined over the last hundred years, as a curriculum problem (1920s-50s), training problem (1960s-80s), learning problem (1980s-2000s), and then a policy problem (1990s-present). During the training problem phase, teaching was seen as a technical transmission enterprise that required the right set of moves to ensure success. Although the process-product research of this period was not particularly interested in student outcomes, the teacher effectiveness research of today can be seen as an extension of that approach from the training problem paradigm. Teacher effectiveness research sits solidly in the policy problem paradigm, which similarly seeks generalizable findings that can be applied to initiatives intended to raise teacher quality. Both of these paradigms rely on quantitative approaches that can provide concrete prescriptions for action, at the level of teacher training or policy, and therefore are drawn to externally measurable indicators of inputs and outputs modeled in ways derived from economics.

In contrast, the learning problem paradigm parallels the interpretive genre and is largely carried out by teacher educators (who tend to be professionalizers) using

qualitative methods (Kennedy, 2008). The learning problem paradigm rejects the technical view of teaching as training and presumes that teaching is an intellectual enterprise that is strongly situated. This research looks at internal teacher characteristics such as attitudes and beliefs or instructional processes (i.e., the portions of the conceptual framework from the Introduction that will not be the focus of this study). Policymakers are less interested in the learning problem paradigm because it does not examine causal questions or make the link to student achievement. The overlap between the learning problem and policy problem paradigms during the 1990s represents a period of struggle between the professionalizers and deregulators. As standards-based reform has come to firmly dominate the field of education in the last decade, so too has the outcomes-focused research paradigm:

> The assumption behind constructing teacher education as a policy problem is that one important way policy makers can meet the challenges of providing a high-quality teaching force is by manipulating those broad aspects of teacher preparation (e.g., teacher tests, subject matter requirements, entry routes) most likely to affect pupil achievement. (Cochran-Smith & Fries, 2008, p. 1082)

Teacher effectiveness research seeks to identify the levers most useful to policymakers by setting up statistical models that operationalize the various inputs available for manipulation. Next, I will look at the basic form of the models used in this type of research.

Basic education production functions to estimate value added by teachers. A function models a mathematical relationship between inputs and outputs, and economists have long used production functions to estimate the output of a firm, industry, or

economy under various conditions. A former economist turned education researcher, Monk (1994) defined an education production function as "shorthand for whatever there is about teaching and learning that is systematic and predictable" (p. 1189). Specifically, all of the different inputs are put on one side of an equation to model the amount of impact they have on predicting the outcome on the other side of the equation. Student achievement is the bottom line metric that is used as the outcome. This type of approach was famously employed in the *Coleman Report* of 1966 (Coleman, et al., 1966) and underlies the basic strategy of most teacher effectiveness studies.

Hanusek and Rivkin (2010a) walked us through the basic education production function:

$$A_g = \theta A_{g-1} + \tau_j + S\lambda + X\gamma + \varepsilon$$

where A_g is the achievement of student A in grade g (the subscript I is suppressed thoughout), A_{g-1} is the prior year student achievement in grade g - 1, S is a vector of school and peer factors, x is a vector of family and neighborhood inputs, $\theta, \lambda,$ and γ are unknown parameters, ε is a stochastic term representing unmeasured influences, and τ_j is a teacher fixed effect that provides a measure of teacher value added for teacher j. (p. 267)

Note that each term on the right hand side of the equation is estimated independent of the other terms; that is, by putting all of the input in the same equation, the model accounts for the relative magnitude of each variable's impact on the outcome. The term of most interest in teacher effectiveness research is τ_j because it represents the amount of variation in student achievement that can be attributed to the teacher, above and beyond the contribution of all the other variables in the model.

The other terms in the equation, often referred to as control variables, intend to capture as many of the known contributors to student achievement as possible so that I can isolate the impact of the teacher. If I don't control for a variable such as the students' socioeconomic status (SES), for example, then teachers of high SES students would likely appear to be significantly more effective than teachers of low SES students. Since we know from research going back to the *Coleman Report* that factors such as the SES of individual students and the SES of their peers matter, I include them in the model so that I can look at what the teacher adds when all of these other factors are accounted for. One of the reasons that teacher effectiveness research has boomed in the past decade is due to the availability of expansive data sets that include the factors needed to build statistical models with appropriate controls.

Perhaps the most important recent advance in accessible data is the availability of longitudinal data that follows students over multiple years, in large part due to the requirement in the No Child Left Behind Act that states establish end-of-year testing in grades 3-8. Longitudinal data allows researchers to factor in prior achievement, either as a control variable as above, or in some cases using gain scores from one year to the next as the dependent variable. The importance of this cannot be understated for teacher effectiveness research. Logically, researchers can get a much better read on the value that a teacher has added to a student's achievement by looking at how much the student gained during the year in that teacher's classroom. Further, many of the unobservables that are not included as variables in a model, such as home influences or internal student traits such as motivation, may be stable and cumulative over time and will therefore be largely captured by and controlled for by incorporating prior achievement.

William Sander is considered a pioneer for developing a widely cited value added model for teacher effectiveness research (Sander & Horn, 1994). Sander, Wright, and Horn (1997), for example, looked at the impact of teachers on student achievement in 54 school districts in Tennessee. The Tennessee state assessment was designed to be vertically scaled so that student gains (current test score minus prior year's test score) could easily be used as the dependent variable. Sander et al. used a statistical model that included measures of intraclassroom heterogeneity and class size to balance out the advantage of teachers whose classrooms were comprised of less academically diverse or fewer students. The study found that teacher effects were the dominant factor and that class size and intraclassroom heterogeneity had little impact: "Effective teachers appear to be effective with students of all achievement levels, regardless of the level of heterogeneity in their classrooms" (p. 63). Notably, however, the study did not account for student SES, meaning that the teacher effect finding could be confounded by systematic variance in SES. For example, if students were grouped into classrooms by their SES level, then what appeared to be differences in student growth between classrooms could have been due to SES rather than teachers. Leaving SES out of the model could cause other variables to exhibit false effects as well. If larger classes tended to have higher SES students, the negative effect of class size may be compensated for by the positive effect of high SES, leading the authors to erroneously assume that class size didn't matter. Although in this case, Sander and Horn (1998) included SES in a follow up study and found that it did not change their conclusions, it is important to recognize that education production function research is limited by the strength of its data and the inclusion of proper controls in the models.

Accounting for non-random sorting and other unobservables. No matter how good the data set, some factors defy measurement and are impossible to include in statistical models. Kennedy, Ahn, and Choi (2008), for example, pointed to teacher self-selection phenomena that cause the assignment of teachers to schools to be nonrandom. Differential migration refers to the fact that teachers tend to seek positions in schools similar to the ones they attended as students, thus matching their background to their students'. Another mechanism involves the tendency (out of necessity) for urban school systems to hire more novice teachers with lower qualifications. These processes "make it difficult to tell whether teachers created their students' achievement levels or whether, instead, students with different achievement levels have attracted different kinds of teachers" (p. 1251). Control variables help account for nonrandom assignment to the extent that students group with teachers in ways captured by the controls (e.g., urban students would likely be low SES as well), but the challenge of including all relevant variables is increased by nonrandom assignment as students group in subtle ways.

Modern teacher effectiveness researchers use fixed effects models to account for unmeasured factors. Betts et al. (2003) provided a detailed explanation of the role of fixed effects in their model. The study tapped a robust data set linking elementary and high school teachers to their students in the San Diego Unified School District. The study was longitudinal, spanning three years, included a number of control variables, and used gain scores. The authors were most interested in modeling the impacts of teacher credentials and did not want unobserved school level characteristics, such as the attitudes of teachers and administrators or school climate, to confound their analysis. By subtracting the mean of a variable at the school level from each observation at a given

school, the authors explained that they were able to "remove all the variations among schools, which leaves only the variation within the school" (p. 122). They fixed student effects in a similar way to account for any unobserved characteristics in students, such as motivation or innate ability, around which students may be grouped into classrooms. In this case, the approach helped ensure that variations in student achievement could be attributed to teacher credentials, even if students with certain types of unobserved characteristics tended to be matched with teachers with certain credentials. The model used the three longitudinal data points for each student to estimate the impact of variables by looking at variations within individual students over time, across different classrooms as they change teachers each year, rather than between students. School fixed effects methodologies are becoming more commonly used to guard against nonrandom sorting between schools, and some studies use student fixed effects to minimize issues of nonrandom sorting between classrooms (Hanushek & Rivkin, 2010a).

Statistical models are by definition simplifications, and it is impossible to fully account for factors such as unobservable variables and measurement error. As noted at the outset and summarized by Hanushek and Rivkin (2010a), the impact of teachers has been established using various data sets with education production functions tweaked to the nuances of the data. The next section considers policy implications of the fact that teachers matter.

Policy considerations. Having established the broad premise that teachers impact student achievement, the analytic framework of value added modeling also allows researchers to attempt to measure the impact of particular teachers. All of the considerations and caveats of the above discussion regarding teacher effectiveness

29

research are amplified when the models are used to try to make attributions at the individual teacher level.

In a recent point/counterpoint exchange, Hill (2009) argued against incorporating individual teachers' value-added scores into moderate- to high-stakes evaluations. She analyzed the empirical evidence from a validity framework standpoint, holding value-added scores to the same standards as any educational assessment, and found that value-added estimates did not exhibit enough stability between years (reliability) to accurately distinguish between teachers unless they were at opposite ends of the effectiveness continuum. Hill also pointed to variables that weren't included in typical value-added models, such as the resources available to teachers and the nonrandom sorting of students between classrooms, and expressed concern that the added incentive of personal stakes may motivate teachers to manipulate student outcomes through test preparation and other adaptation strategies. She urged the reader to consider a revised interpretation of value-added scores:

> Rather than assuming a value-added score is an indicator of teacher quality or effectiveness, as is often done in current debates, we must more accurately characterize these scores as representing not only teacher quality but also bias due to student selection, the effect of other resources on student achievement, and a generous amount of measurement error. (p. 706)

Based on this interpretation, she argued against the use of value-added scores as any part of teacher evaluations.

Hill's sparring partner conceded that there are many unknowns but that the potential benefits of extending accountability to the individual teacher level merit giving

value-added measures a chance: "A strong argument can be made for facilitating voluntary experimentation with, and rigorous evaluation of, accountability systems that include teacher value-added as one component" (Harris, 2009, p. 710). Although this seems like a somewhat cautious endorsement from the pro-value-added side of the debate, it echoes the qualified advocacy of other researchers who agree with the idea of using value-added scores to inform teacher evaluation but are unconvinced that it is ready for prime time as the sole determinant of individual teacher effectiveness (Hanushek & Rivkin, 2010a; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Odden, Borman, & Fermanich, 2004; Rivkin, 2009; Rothstein, 2010). In any case it would appear that value-added measures are destined to become part of the landscape of teacher evaluation with substantial funding support from the federal government (for example, see Patrick, et al., 2010) and initiatives such as the Gates Foundation's *Measures of Effective Teaching Project* (Measures of Effective Teaching Project, 2010). In Massachusetts, a new educator evaluation policy requires student growth to be used for part of teacher evaluation starting in school year 2014-15.

With less than 1% of teachers rated unsatisfactory using current methods, teacher evaluation has much room for improvement (Weisberg, Sexton, Mulhern, & Keeling, 2009). Yet if education does indeed "face a human capital crisis" (Goldhaber & Hannaway, 2009, p. 12), reformers must look to strategies for improving the current workforce beyond changes to teacher evaluation. Although districts and schools largely rely on professional development to improve teacher effectiveness, the research base linking professional development to student achievement is thin and mixed, with some evidence supporting content-based professional development (Harris & Sass, 2007) and

other evidence that moderate investments produce no effect (Jacob & Lefgren, 2004). Although there may be perceptions by some observers that the view that good teachers are born and not made is receding (Green, 2010), the fact that teachers vary in effectiveness has not yet been complemented by an understanding of how to improve effectiveness.

From a policy standpoint, the alternative to building capacity of the current workforce is to alter the make-up of the workforce by changing the personnel. Which brings us back to the deregulation versus professionalization power struggle. Deregulators argue that value-added methods are robust enough to at least incorporate into teacher evaluation as a way of identifying chronically underperforming teachers and removing them from the workforce. Hanushek (2009) calculated that "deselecting" the bottom 10% of teachers and replacing them with average teachers would raise student achievement nationally by a substantial 0.5 standard deviations. On the selection side, deregulators advocate minimizing entry requirements and leaving hiring decisions to local discretion. While there is evidence that principals can identify and hire effective teachers (Jacob & Lefgren, 2008), the state has an obligation to ensure that teachers who are part of the labor market are at least minimally competent. States use teacher licensure policies to perform this gatekeeping function, based on markers that are deemed to be predictors of future teacher effectiveness. In the next section I examine what the teacher effectiveness literature has to say about which teacher characteristics predict teacher effectiveness.

Predicting Teacher Effectiveness: Linking Specific

Teacher Inputs to Student Achievement

As described in the last section, the basic education production function includes a term for teacher effects plus a number of control variables related to student background and school context. By breaking up the teacher term into multiple factors that each represents a different characteristic, this same approach can be used to parse out the impacts of specific teacher inputs. This section first looks at the most recent relevant literature reviews in this area, then reviews the findings of education production function studies that involve teacher knowledge characteristics most relevant to this study, and ends with a brief examination of other characteristics prominent in the literature, including licensure status, cognitive ability, teacher race, teacher gender, and experience. My conclusion is that although there is evidence that teacher licensure tests and teacher knowledge impact the mathematics achievement of elementary students, the overall evidence of the effects of teacher inputs is quite mixed and the predictive power of these and other factors is likely to depend on context.

This review focuses on exemplary teacher effectiveness research by concentrating on quantitative studies that exhibit the features of value-added models, including statistical methods with proper controls for student background, inclusion of prior achievement to model student achievement gains rather than status, and large data sets that link students (and their outcomes) directly to teachers. Due to these constraints, this review is quite limited. Most of the studies in the broader field of teacher education use qualitative methods (Kennedy, 2008), but they do not qualify as teacher effectiveness studies as defined here. As noted previously, the current policy problem paradigm in

teacher education research privileges the teacher effects genre, and researchers who are not teacher educators dominate teacher effectiveness research (Wilson, Floden, & Ferrini-Mundy, 2002). This literature review is intended to inform this study of the impacts on student achievement of a teacher licensure test, and so teacher effectiveness studies involving teacher characteristics available to the state along with or instead of teacher test scores are most relevant.

Selected Literature Reviews

This literature review will build off of two existing reviews of teacher effectiveness research (Boyd, et al., 2007; Wayne & Youngs, 2003). These reviews taken together suggest fairly strong evidence for the importance of content knowledge for mathematics teachers, especially at the high school level, with mixed results for other characteristics. These findings are corroborated by reviews and meta-analyses that go beyond teacher effectiveness research to include studies that define teacher competence in ways other than impacts on student achievement and include qualitative methods (Cochran-Smith & Zeichner, 2005; D'Agostino & Powers, 2009; Goe, 2007; D. N. Harris & Rutledge, 2010; Kennedy, 2008; Rice, 2003; Wilson, et al., 2002).

Major teacher effectiveness literature review yielded mixed results. Wayne and Youngs (2003) conducted a thorough review of studies of the relationship of student achievement gains and teacher characteristics with the intent of interpreting the research base for policymakers. They looked at 21 studies with data collected in the United States that met their criteria of measuring student achievement on standardized tests and using value-added approaches, which they defined as accounting for prior achievement and SES. They consciously avoided the common literature review methods of either tallying

34

results or formal statistical meta-analysis, arguing that these approaches obscure the relative strengths and weaknesses of individual studies. Instead, for each category of teacher characteristic, the authors interpreted each individual study, synthesized the results, and then commented on implications, so that policymakers received the full benefit of their insights.

For the category of teacher test scores, Wayne and Youngs split the studies into three groups: (a) studies involving teacher licensure examination scores, which are most relevant here, (b) subsequent student achievement studies involving tests of teachers' verbal skills, and (c) studies involving other test score measures. Only two studies were in the first category of licensure exam scores. One looked at the relationship of National Teachers Exam scores and outcomes for student in Philadelphia, included a variety of statistical controls, and found no relationship for secondary students and a negative relationship at the elementary level (Summers & Wolfe, 1975). The other study found a relationship using data from Texas, but the analyses were at the school district level (Ferguson, 1991). Five studies from the other two groups produced similarly mixed findings. Wayne and Youngs noted that the two negative findings among the seven studies both controlled for college ratings and summarized their synthesis with the statement that "Test scores matter if college ratings have not already been taken into account" (p. 100). This seems like a strong endorsement for the power of licensure test scores based on thin evidence; a more supportable conclusion would be that test scores have no impact in a model that appropriately controls for college selectivity. It should be made clear to policymakers that teacher test scores may not provide any information that is not already available from the rating of a candidate's college. Ultimately, Wayne and

35

Youngs did not give a wholehearted endorsement to teacher licensure testing, recommending the expansion of performance assessments and calling for predictive validity research on teacher licensure exams.

The evidence for the college rating category suggested some relationship between college ratings and student achievement, though one of the three studies was indeterminate and the other two found no relationship for subcategories of students (i.e., certain grades or races). As mentioned above, it is unclear whether college selectivity is a factor unto itself or whether it signals an underlying trait such as cognitive ability (which could explain its confounding with test scores).

For the degrees and coursework category, Wayne and Youngs reported convincing findings for the impact of degrees and coursework in mathematics at the secondary level. A series of studies utilized data from the NELS: 88 data set, a nationally representative survey of 24,000 grade 8 students with subsequent data collections when they were in 10th and 12th grades. This was truly a unique data set at the time of the Wayne and Youngs review, and great weight was given to studies that tapped into this large-scale longitudinal data base that included various teacher and student characteristics. Wayne and Youngs pointed to two studies by the same authors as evidence for the importance of mathematics degrees. Goldhaber and Brewer (1997) ran separate models for 10th grade student achievement and found that the model that included variables for subject matter produced effects for undergraduate and graduate degrees in mathematics. Goldhaber and Brewer (2000) had nearly identical findings for 12th grade students. The contributions of mathematics-specific degrees in these studies were established with robust models that included controls for student background,

certification, and teaching experience. These same models did not find significant effects

for degrees in other subjects. Similarly, using another longitudinal nationally

representative data set, the Longitudinal Study of American Youth, Monk and King

(1994) found positive effects of coursework in mathematics and none in science.

Another study found no relationship between the number of mathematics courses and

elementary student achievement. Therefore, Wayne and Youngs (2003) concluded that

degrees and coursework made a difference only for mathematics.

The teacher licenses category also exhibited subject-specific results in

mathematics only. Wayne and Youngs based their conclusion on the two Goldhaber and

Brewer studies analyzed in the degrees and coursework category. Goldhaber and Brewer

(2000) is particularly relevant because it focused on certification of 12th grade teachers.

The study found that after controlling for degrees, there was no significant difference in

terms of student achievement between teachers who held standard certification and those

who indicated their certification status as emergency (who are typically alternatively

certified teachers). Goldhaber and Brewer concluded that "This result should, at the very

least, cast doubt on the claims of the educational establishment that standard certification

should be required of all teachers" (p. 141).

The study did indeed provoke the educational establishment. In a classic

exchange between professionalizers and deregulators, Darling-Hammond, Berry, and

Thoreson (2001) critiqued the Goldhaber and Brewer (2000) findings methodologically,

provided a literature review that was intended to show that the research base in support of

certification was strong, and generally accused Goldhaber and Brewer of overstating the

import of their study. In a rejoinder, Goldhaber and Brewer emphasized their statement

from the original study that "Our study does not definitively answer the important policy question of whether imposing more rigorous standards in teacher licensure will lead to better student achievement" (Goldhaber & Brewer, 2001, p. 141). They argued that the methodological discussion of Darling-Hammond et al. (2001) was off-base and the literature review included a lot of weak research. They agreed with Darling-Hammond et al. that more research was required, but insisted that policymakers demand research in the teacher effectiveness vein.

The Wayne and Youngs (2003) literature review is worthy of detailed examination because it exemplifies various aspects of teacher effectiveness research. First, good studies were extremely hard to find before the turn of the century. Researchers are increasingly gaining access to state- or district-wide administrative databases that include the extensive information that used to only be found in research-specific data sets, as we will see in the next subsection. Second, no teacher characteristics stand out as clearly predicting teacher effectiveness in all cases. Mathematics content knowledge seems important at the secondary level, but findings even in that area are contentious. And finally, the mixed nature of findings justifies calls for more studies to replicate results and further examine the links between specific teacher inputs and student achievement in various contexts.

Recent review including longitudinal studies produces more mixed findings. A more recent review covers much of the same territory as Wayne and Youngs (2003) and also analyzes another five or so years of studies. Boyd, Goldhaber, Lankford, and Wyckoff (2007) similarly focused on providing insight to policy makers, specifically honing in on the alternative certification question that formed the backdrop of the debate

sparked by Goldhaber and Brewer (2000). After examining studies about teacher

characteristics on student achievement that produced mixed results, as well as

indeterminate evidence regarding the effect of certification requirements on the pool of

teacher applicants, Boyd et al. concluded that "the research evidence is simply too thin to

have serious implications for policy" (p. 45) and urged more research.

Specific to teacher licensure exams, based on three studies with "strong research

designs and good data," Boyd et al. summarized that

> In both North Carolina and New York City, these studies find, performance on
>
> required certification exams is predictive of teachers' abilities to increase student
>
> achievement, especially in mathematics, but exam scores affect student
>
> achievement less than, for example, teacher experience does. Thus the exams do
>
> distinguish among teachers, but only relatively weakly. (p. 59)

Since these studies used multiple control variables, teacher scores showed up as one of

many effects within a package of qualifications. For example, the New York City study,

unpublished at the time of Boyd et al.'s review, found that increased qualifications of 4th

and 5th grade teachers in the poorest schools in New York City accounted for substantial

portions of closing achievement gaps. The results were weaker at the middle school level

and completely insignificant across all qualifications for English Language Arts.

The two studies in North Carolina referenced by Boyd et al. tapped into an

extraordinary statewide database. Clotfelter, Ladd, and Vigdor (2006) mentioned that at

the time of the study, it was the only known state level data set that went to the classroom

level. Also, North Carolina has a very prescriptive statewide course of study, and

therefore the curricula taught by teachers at the classroom level is likely to align better

with the state assessment than may be the case in other states. The data set is somewhat flawed by the fact that students are linked only to the teacher who proctored their exam, but it is a fair assumption that in the great majority of cases these are the classroom teachers. The North Carolina data continues to be used and recent studies have taken steps to minimize false matches (for example, see Goldhaber & Hansen, 2009).

Clotfelter et al. (2006) used a cross-section of the data set and examined 3,842 teachers who taught fifth grade students during the 2000-2001 school year. Since this non-longitudinal data did not allow them to utilize student fixed effects, the authors mitigated against within school nonrandom sorting by limiting their analysis to schools where student achievement was distributed evenly across classrooms. The authors found that experience impacted student achievement in both mathematics and reading, while the effects of teacher licensure exam scores only surfaced in mathematics, and college coursework and National Board Certification were insignificant. As noted by Boyd et al., however, the effects of the licensure exam were quite small, with a one-standard-deviation increase in licensure exam performance predicted to increase fifth grade students' performance by .01 of a standard deviation, compared to an increase of .10 of a standard deviation for every one-standard-deviation increase in experience.

Goldhaber (2007) used ten years of data (1994-1995 through 2004-2005) for North Carolina students in grade 3-6. The longitudinal nature of the data allowed him to use fixed effects to mitigate against nonrandom sorting. Within the ten years covered by the data set, North Carolina had raised its cutoff score, allowing Goldhaber to examine the effectiveness of teachers who had received licenses under less stringent testing requirements. Also, North Carolina uses the Praxis licensure exams that are also used in

other states, so Goldhaber modeled the consequences of implementing the higher cutoffs used by Connecticut. Noting that the fixed effects model indicated that the predictive power of licensure tests was fairly small, Goldhaber pointed to a policy tradeoff:

> Despite the testing, many teachers whom we might wish were not in the teacher work force based on their contribution toward student achievement are nevertheless eligible because they scored well on the test. Conversely, many individuals who would be effective teachers are ineligible due to their poor test performance. For example, the results suggest that upping the elementary teacher licensure lest standard from the one currently used in North Carolina to the higher standard used in Connecticut would lead to the exclusion of just 0.2 percent of the teacher work force who are estimated to be very ineffective teachers, but would also result in the exclusion of 7 percent of the teacher work force who are estimated to be effective. (p. 767)

The quantitative expression of this tradeoff serves notice to policymakers that predictive validity of licensure exams is a necessity and that the tests can only function as appropriate screening mechanisms to the extent that the association with student achievement is strong.

Although administrative data sets such as those available in New York and North Carolina provide researchers with opportunities to explore teacher effectiveness in much more sophisticated ways than the studies examined in Wayne and Youngs (2003), Boyd et al. (2008) were not convinced that the evidence base for teacher licensure exams, much less teacher preparation, provided useful guidance for policymaking. The theme that continued through both of these major literature reviews was that teacher

characteristics, especially those related to teacher knowledge (i.e., exam scores and degrees), produced stronger effects in mathematics than other subjects. The next section looks more closely at the literature related to teacher knowledge in mathematics.

Teacher Knowledge

This section looks more closely at the impact of teacher knowledge for mathematics, first by looking more closely at the use of proxy measures for teacher knowledge, which have produced mixed results with a fair amount of evidence that math-specific degrees and coursework benefit the mathematics achievement of high school students. Then I examine recent studies that find consistent effects of licensure test scores on student growth in mathematics. Finally, I examine a research agenda focused on the mathematical knowledge for teaching elementary mathematics that points to the importance of attending to pedagogical content knowledge in addition to subject matter content knowledge.

Proxies for teacher knowledge in mathematics show mixed results. Although assessments like teacher licensure exams may be the most direct way of measuring teacher knowledge, data about the educational background of teachers is ubiquitous and many studies have used markers such as degrees and majors as proxies for subject matter knowledge. Findings in this area are somewhat mixed but tend to point to proxies for mathematics knowledge as predictors for student achievement in mathematics at the high school level. Studies that use proxies suffer from the methodological weakness that teachers self-select these characteristics and so it becomes difficult to disentangle the markers from underlying motivation and other unmeasured traits that could confound the role of these proxies as measures of teacher knowledge.

Kennedy, Ahn, and Choi (2008) reviewed 19 studies to understand the relative impacts on mathematics achievement of content knowledge, pedagogical knowledge, and pedagogical content knowledge. The authors treated degrees and coursework in mathematics as proxies for content knowledge, in education as pedagogical knowledge, and in mathematics education as pedagogical content knowledge. They limited their review to high quality studies that linked students to teachers, used sophisticated statistical models, and modeled student achievement gains. They normalized the effects reported in each study to represent the percent of average annual gains and visually represented each study as a hash mark to look for patterns.

The authors performed two within-study comparisons. The first examined Monk (1994), which was excluded from Wayne and Youngs (2003) because of inadequate controls for SES. Monk found benefits for mathematics education course taking and smaller benefits for mathematics courses, with diminishing returns. Monk also found a surprising negative effect of majoring in mathematics. Kennedy et al. (2008) used this finding to point to the issue of unmeasured variables: "these measures of course-taking reflect *both* the knowledge gained from the courses *and* the teachers' original interests and motivations that motivated her to take the courses in the first place" (p. 1257). This is an example of the overall weakness of any studies, including almost all teacher effectiveness research, that use proxies for which teachers can self-select. The second within-study comparison by Kennedy et al. involved the Harris and Sass (2006) study, which broke out results between elementary and secondary teachers. Kennedy et al.'s analysis of Harris and Sass showed benefits of education and mathematics education courses at the elementary level, with no clear benefits for mathematics courses at the

43

elementary level or mathematics education courses at the secondary level and negative

effects of both education and mathematics courses at the secondary level. The results of

these two highlighted studies, therefore, yielded mixed results for the effects of degrees

and coursework.

The full analysis of all 19 studies supported the within-study analyses by showing

benefits of course taking within each of the three domains and at both the elementary and

the secondary level, with mixed evidence for majoring in mathematics or education as an

undergraduate. There were clear differences between domains only for advanced degrees

at the secondary level, with advanced mathematics degrees showing clear benefits and

advanced education degrees producing no effect in one study and a negative effect in

another.

Kennedy et al. concluded that coursework in each domain seemed to be helpful to

a point, with diminishing returns causing mixed results for whether majoring in

mathematics is beneficial. This study has been criticized for giving equal weight to

various studies regardless of methodologies or the specific controls used in their

statistical models (Darling-Hammond, 2008). Other literature reviews have been more

positive about the benefits of math-specific majors and advanced degrees for student

achievement in mathematics at the high school level (Cochran-Smith & Zeichner, 2005;

Goe, 2007; Rice, 2003; Wayne & Youngs, 2003), yet recent studies continue to show no

or negative effects of mathematics degrees for elementary students' mathematics

achievement (Buddin & Zamarro, 2009; Croninger, et al., 2007). It seems fair to

conclude that teachers' knowledge is an important determinant of effectiveness,

especially at the secondary level, and that direct measures of knowledge have the potential to predict effectiveness more accurately than proxies.

Licensure test scores show mixed results, though stronger in math. Since Boyd et al. (2007) reviewed the literature, there have been four studies published that look at the predictive validity of teacher licensure tests for student achievement. These studies show increased attention to expressing effect sizes in ways that would be relevant to policymakers and also attend to a greater extent than previous teacher effectiveness research to equity concerns. Three of these studies used statewide data from North Carolina and produced effect sizes of teacher licensure test scores on student math performance that range from .01-.05 standard deviation units, while the other study uses a different data set and found no association.

Licensure test scores impact as one component of teacher credentials. Clotfelter, Ladd, and Vigdor (2007) tapped into the longitudinal data set from North Carolina and examined the impact of teacher credentials for grade 3-5 students in years 1995-2004. The study found that teacher test scores were much stronger predictors of students' mathematics achievement than their reading achievement. The authors also found that the effects of test scores in mathematics were nonlinear. Their linear model suggested that a one-standard-deviation difference in teacher test score translated into a .015 standard deviation increase in student achievement; thus comparing teachers at either end of the distribution who were four standard deviations apart yielded a prediction that the student achievement of these teachers would differ by .060 standard deviations. When the test score continuum was segmented, however, the overall difference between teachers at the two extremes was 0.13, over twice as large as the linear model.

When taken as a package, the effects of the qualifications are more practically significant. The authors modeled the difference between teachers with weak credentials (across experience, college selectivity, licensure status, license test score, graduate degree, and National Board Certified) versus teachers with strong credentials and found that a one-standard-deviation difference in overall weakness produced a 0.21 difference in student achievement for mathematics and 0.12 difference for reading. The authors noted that the effect of the credentials package is comparable to the effect of a five-student change in elementary class size. They also drew on the stable estimates of overall teacher impacts to develop scenarios that suggested that teacher credentials would account for a sizeable portion (e.g., 33%-66% in one scenario) of the total variation in teacher effectiveness. These types of comparisons are important to put the impact of teacher credentials in context relative to variations in teacher effectiveness that are attributable to unmeasured characteristics and practices.

Clotfelter, Ladd, and Vigdor (2010) used a data set from North Carolina at the secondary level to look at the association between teacher credentials and the achievement of high school students. They analyzed four cohorts of tenth graders (from 1999-2003) using results on the statewide end-of-course assessments. The study is noteworthy because although the data was cross-sectional, the authors used student fixed effects by looking at variations within students across subjects, rather than over time. As with the typically used longitudinal cross-classroom fixed effects, the cross-subject fixed effects approach allowed them to statistically minimize nonrandom sorting within schools.

Once again, teacher test scores in mathematics produced a significant effect, with a one-standard-deviation difference in teacher score predicting a difference in student scores on either algebra or geometry of 0.047 standard deviations. In contrast, other subjects produced smaller effects (0.016 in biology), insignificant effects (political science), or negative effects (-0.021 in English). The authors performed an analysis of the credentials package (certification by subject area, licensure test score, National Board Certification, and experience), similar to Clotfelter et al. (2007), and found a comparable effect size of .23 standard deviations.

The authors devoted considerable analyses to equity issues. They reported evidence of uneven distribution of teacher credentials and showed that students who were poor and nonwhite were more likely to be taught by less effective teachers. Looking at matching effects in Algebra I specifically, the study found that male teachers negatively impacted the achievement of female students (-0.1 standard deviations) and black teachers produced negative effects on the achievement of white students (-0.08 standard deviations). These effect sizes were especially troubling in comparison to the 0.047 reported for teacher test scores. The authors concluded that these findings "should be cause for serious policy concern" (p. 679) and urged further research and attention to these issues.

Licensure test score impact as a research focus. Goldhaber and Hansen (2009) used a North Carolina data set spanning 11 years that included 175,000 students in grades 4-6, taught by 4000 teachers. The authors brought a strong policy focus to the study by concentrating on the predictive validity of the licensure exams based on different policy functions of the exams and the background characteristics of teachers.

First, Goldhaber and Hansen looked at the screening function of the testing program by calculating the association between passing the test and student achievement. There were two pools of teachers that were compared to teachers who had passed the current screen: (1) unlicensed teachers who were teaching on waivers and (2) teachers who were licensed before a policy change in 2000 who would not have passed the screen under the new, higher cut-score. In reading, the students of teachers who had passed the current screen did not perform any better than the students of teachers in the comparison group. For mathematics, however, there was a positive and significant value to the screen (at the .05 alpha level), with a passing score on the screen predicting a .05 standard deviation increase in student growth. The authors also established that the screen did not vary by the demographic background of the teachers.

Next, Goldhaber and Hansen went beyond cut-scores to focus on the signaling function of the testing program by modeling the impact of teacher scores on student achievement. Whereas for the screening models, a combined score was used to determine whether a teacher passed or failed, the signaling model looked at each of the two required tests separately. The Praxis II Curriculum, Instruction, and Assessment (CIA) test showed stronger association on student achievement, with statistically significant results in both math and reading (at the .1 alpha level). The Praxis II Content Area Exercises (CAE) test, in contrast, only showed significance in mathematics (at the .1 alpha level). In terms of effect size, a one standard deviation increase in teacher test score was associated with an increase in student gains by of about .01 standard deviations.

Goldhaber and Hansen also found systematic variations in the predictive power of the licensure tests by the demographic background of teachers, with the CIA serving as a significantly better predictor of effectiveness for white teachers than the pooled model in both reading and mathematics (at the .1 and .05 alpha levels, respectively), and predicting the effectiveness of female teachers better than male teachers for both reading and mathematics (at the .1 and .05 alpha levels, respectively). Furthermore, CAE results predicted the effectiveness of black teachers significantly better than teachers of any other race (at the .1 alpha level). The authors argued that although for state licensure purposes the tests are simply pass-fail, local districts and schools could theoretically use test scores to inform hiring decisions, and so the differential predictive validity by teacher background was cause for concern.

Overall, Goldhaber and Hansen (2009) had a few methodological weaknesses worth mentioning. First, as mentioned previously, the teacher-student link was based on the teachers who monitored the students' exams. Although there was no way to validate that these teachers were indeed assigned to the students with whom they were linked, the authors excluded students taught by multiple teachers and also noted that the effect sizes they found for common variables were comparable to studies with direct teacher-student linked data. Second, the statistical model for screening did not control for teacher variables with the rationale that testing policies are blind to other teacher attributes. It would have been useful to examine models that included the other teacher variables as well, however, in order to compare the effect of teachers' test scores to the other attributes in a way that would allow an estimation of the usefulness of the information provided by test scores above and beyond the other, readily available information.

Without teacher control variables included in the model, it is possible that a characteristic such as advanced degrees could have predicted student achievement better than the licensure test score. Finally, the study reports many effects as statistically significant at an alpha level of .1 rather than the more commonly accepted alpha level of .05. At the more stringent threshold for significance, the association between the mathematics screen and student achievement is the only finding that holds up for all teachers, along with some of the findings for the differential signaling function based on teacher background.

Licensure test score lack of impact in a different data set. Buddin and Zamarro (2009) used a data set of 2,738 Los Angeles Unified School District (LAUSD) teachers who graduated from the California State University (CSU) system in the years 2000-2006. The data was compiled by the CSU system and since most teachers in LAUSD either received their licenses before 2000 or did not attend CSU, the data set represented only about 17% of LAUSD teachers (38% of teachers in their first three years of teaching). Although it is comprised of a subset of teachers who are not necessarily representative, the data set has an advantage over the North Carolina data because it matches students directly to teachers rather than assuming that teachers who proctor students' exams are also their classroom teachers.

Using both state and teacher fixed effects, the authors looked at the impacts of the three licensure exams required for California teachers: the California Basic Educational Skills Test, required for admission into a teacher preparation program; the California Subject Examinations for Teachers, the elementary teacher version of which covers multiple subjects; and the Reading Instruction Competence Assessment, required for all elementary teachers. Whether taken separately or combined, teacher test scores did not

predict student achievement. The authors looked at nonlinearity and various other possibilities, but all models showed no impact of test scores on teacher effectiveness. Findings were similar for the other credentials that were analyzed, with advanced degrees showing no effects and experience very weakly related to student achievement.

A major weakness of the study is that the authors had no way of looking at the effectiveness of teachers who hadn't passed the tests and received their licenses. The North Carolina data set had the benefit of cut-scores that had changed over time, which allowed those studies to examine teachers who had come into the system under different standards. The North Carolina method is imperfect as well since it does not compare teacher effectiveness for teachers who enter the teaching force at the same time with different passing results. This study has the advantage of the temporary conditional cut score which sets up a natural quasi-experiment whereby I will be able to compare the effectiveness of teachers who exceeded the official cut-score to teachers who scored in the conditional range and are temporarily allowed to teach. Also, the Massachusetts data links teachers directly to students. Although teacher licensure tests have consistently shown predictive power, especially in elementary mathematics, there have been relatively few studies and they are based on only three data sets (North Carolina, New York City, and L.A.). The Massachusetts subtest is also unique because it is designed to assess both content knowledge and pedagogical content knowledge, the importance of which has been established in the literature discussed in the next section.

Mathematical knowledge for teaching solidly supported. Lee Shulman is often credited with bringing the idea of pedagogical content knowledge to the field of education (Shulman, 1986), and Deborah Ball and her associates at the University of

Michigan have taken Shulman's work and applied it to elementary mathematics. Hill, Rowan, and Ball (2005) described mathematical knowledge for teaching (MKT) as

the mathematical knowledge used to carry out the *work of teaching mathematics*. Examples of this "work of teaching" include explaining terms and concepts to students, interpreting students' statements and solutions, judging and correcting textbook treatments of particular topics, using representations accurately in the classroom, and providing students with examples of mathematical concepts, algorithms and proofs. (p. 373)

MKT is distinct from content knowledge, which includes mathematical facts, concepts, and procedures, as well as pedagogical knowledge about how to teach mathematics. It is not exactly subject matter or pedagogy, but the combination of the two as applied to student thinking. Hill et al. argued that production function studies may be turning up mixed results on teacher tests because the assessments used in those studies did not address MKT.

After developing a valid and reliable multiple-choice measure of MKT (H. C. Hill, Schilling, & Ball, 2004), Hill et al. (2005) used the instrument as part of an evaluation of school improvement efforts under the Comprehensive School Reform federal grant. The study included 334 first-grade teachers and 365 third-grade teachers from 115 elementary schools in 15 states. The data collection was massive and included information on teacher and student background as well as school and classroom contexts. Student outcomes were measured using the Terra Nova assessment and the scale scores were used to calculate gains in a value-added framework. The main statistician on the research team, Brian Rowan, was the lead author of another study evaluating the nuances

52

of various approaches to education production function research using large data sets (Rowan, Correnti, & Miller, 2002), so the statistical models were of the highest quality.

Hill et al. (2005) found that teachers' MKT scores were strong predictors of their students' achievement in mathematics. A one-standard-deviation difference in teachers' MKT score translated to a 2.25-point gain on the Terra Nova, interpreted by the authors as two to three weeks of extra instruction in a school year compared to an average teacher. The effect sizes for MKT were far larger than any of the other teacher characteristics in the model, including certification, course taking, and experience. Moreover, the effect size for MKT was comparable to SES, allowing Ball to argue elsewhere from a social justice standpoint that increasing teachers' MKT would be a solid strategy for preventing achievement gaps (Ball, Hill, & Bass, 2005).

The association between MKT and student achievement have been replicated in other contexts (Baumert, et al., 2010), and states such as Massachusetts advocate the use of MKT assessments to evaluate professional development. The research agenda of Ball and her colleagues has turned to exploring how exactly MKT plays out instructionally in order to effect student achievement (Charalambous, 2010; Heather C. Hill, et al., 2008). Although the MKT assessments are not validated to support conclusions about individual teachers, this body of research led Massachusetts to incorporate items intended to measure MKT in the MTEL mathematics subtest. This study will see how an assessment with these types of items performs in the context of teacher licensure.

Other Paper Qualifications and Characteristics

The package of teacher characteristics used in teacher effectiveness studies varies somewhat from study to study due to the data that is available. In addition to

indicators of teacher knowledge (advanced degrees, mathematics majors) and student background variables (e.g., prior achievement, SES, and race), researchers typically develop statistical models with variables that include some combination of licensure status, cognitive ability, teacher race, teacher gender, and experience. All of these characteristics have mixed results except for experience, which is consistently supported as being an important factor up to 3-5 years.

Licensure status. Assessing the effects on student achievement of whether a teacher is licensed or not, or what kind of license they have, poses a challenge for researchers for a variety of reasons. First, the definition of licensure varies substantially by state, grade level, and subject. Second, licensure often encompasses other important variables such as college degree and the effects of licensure can be difficult to disentangle. Third, in some cases licensure requirements set such a low bar that there are few teachers who don't meet the minimal standard (for example, regarding NCLB Highly Qualified status, see Hanushek & Rivkin, 2010b). For these and other reasons, the evidence continues to be mixed, with some recent teacher effectiveness studies showing benefits of licensure (Boyd, et al., 2008; Clotfelter, et al., 2007, 2010; Neild, et al., 2009) and other studies, at the elementary level, showing no impact (Jepsen, 2005) or a negative association between standard licensure and student achievement (Betts, et al., 2003; Phillips, 2010). The recent literature, therefore, continue to support the conclusions of previous reviewers who found the evidence for licensure to be most convincing for mathematics at the high school level (Cochran-Smith & Zeichner, 2005; Goe, 2007; Rice, 2003; Wayne & Youngs, 2003).

Cognitive ability. From the time of a Carnegie Foundation study in the 1930s that found teachers to have lower cognitive abilities than the general population, it has been widely believed that the teaching profession does not draw the most talented professionals (Sedlak, 2008). Although there is evidence that as other professions provided more employment opportunities for women, the cognitive skills of the teacher pool have steadily declined (Corcoran, 2009), there is more positive news recently for educators in studies that showed that less talented people leave teaching at each hurdle and that the SAT scores of teachers who complete licensure programs are on par with other professions (Cochran-Smith & Zeichner, 2005; Zumwalt & Craig, 2008). While college selectivity has shown some effects when used as a proxy for cognitive ability (Rice, 2003; Rivkin, Hanushek, & Kain, 2005; Wayne & Youngs, 2003), recent studies have found an association between student achievement and slightly more direct indicators of cognitive ability such as undergraduate grade point average (Kukla-Acevedo, 2009), SAT scores (Boyd, et al., 2008), and preparatory program grade point average (D'Agostino & Powers, 2009).

Race and gender. In their study of the demographics of the teaching force and impacts on outcomes, Zumwalt and Craig (Zumwalt & Craig, 2008) reported that the research literature is mixed about the effects of teachers' race and gender on student achievement. Although some studies report no association, Zumwalt and Craig noted that the statistical models may not include adequate controls for student and teacher SES. More recently, Munoz and Chang (Munoz & Chang, 2008) found no association between teacher race and student achievement in high school reading. Perhaps the strongest evidence for the impact of race comes from Dee (2001), which used an experimental

design and found increased student achievement of black students when taught by black teachers compared to white teachers. The recent studies involving teacher test scores in North Carolina also showed effects based on the matching of student and teacher race, and to a lesser extent gender (Clotfelter, et al., 2010; D. Goldhaber & Hansen, 2009).

  Experience. Although Wayne and Youngs (2003) did not include experience as a category, other literature reviews concluded that experience within the first 3-5 years was positively associated with student achievement, after which it showed no effect (Harris & Rutledge, 2010; Rice, 2003). Kukla-Acavado (2009) found that experience erased the initial advantage of undergraduate grade point average within three years, suggesting that it could have a leveling effect. Some recent studies reported that experience may matter less at the elementary level than the secondary level (Buddin & Zamarro, 2009; Jepsen, 2005) and less in math than reading (Croninger, et al., 2007; Rockoff, 2004).

Teacher Licensure Tests as Measures of Teacher Effectiveness

  Teacher licensure testing policies have arisen within the context of standards-based education reform and a corresponding shift in the balance of power from professionalizers to deregulators. The new Massachusetts mathematics subtest is in line with deregulation because it emphasizes the measurement of teacher knowledge rather than building professional practice through pre- and in-service professional development. Although the Massachusetts test exhibits surface features of content validity, the research base shows that licensure tests do not have a good track record of establishing empirically-based validity. The historical context that situates the mathematics subtest and persistent question marks about teacher licensure testing generally suggest that a

predictive validity study of the Massachusetts mathematics subtest is of great interest for researchers and useful for policymakers.

Historical context of teacher licensure testing

Since the early 2000's, nearly every state has required prospective teachers to pass licensure tests to earn their teaching license (Rotherham & Mead, 2004). Although deregulators thus seem to dominate the debate at the moment, the professionalization movement seemed poised for greater influence as recently as the mid-1990s. Imig and Imig (2008) traced the abrupt shift in the balance of power by highlighting the central role played by quasi-non-governmental organizations (quangos), private organizations that have government involvement and support. On the professionalization side, one notable quango is the National Board for the Professional Teaching Standards (NBPTS), which developed its own standards and certification system and has garnered over $100 million in federal appropriations. The 1992 reauthorization of the Higher Education Act of 1965 recognized NBPTS as a certifier of the best teachers while strengthening the role of state agencies to set certification standards. In 1994, a powerful quango emerged as Linda Darling-Hammond led the formation of the National Commission on Teaching and America's Future (NCTAF), funded by the Carnegie Corporation and the Rockefeller Foundation. NCTAF commissioned testimony from prominent scholars and built upon prior efforts to issue the professionalization manifesto *What Matters Most: Teaching for America's Future,* calling for high quality teacher preparation, autonomous professional standards boards, performance-based teacher assessment, and improved inservice professional development. The report resulted in federal investments of $30 million and

unprecedented philanthropic grant support for its implementation along with substantial interest from policymakers.

NCTAF had advanced the shift of authority from local campuses to national entities and numerous Congressional proposals were prepared in anticipation that the Higher Education Act of 1998 would create a national system of certification. Organized opposition from the deregulation camp, however, intervened to ensure that the shift to centralized authority worked in their favor. A group of think-tanks and other external organizations arose in opposition to the educational establishment (for an analysis of the organizational dynamics in teacher education, see Wilson & Tamir, 2008) and made their message felt by "challenging the empirical legitimacy of each plank of the professionalization platform" (Imig & Imig, 2008, p. 894). The Fordham Foundation made a pitch to centrists with the report *Better Teachers, Better Schools,* which attacked NCTAF and called for more local discretion in teacher preparation, certification and hiring. The Fordham Foundation warned that "The regulatory approach is also bound . . . to undermine the standards-and-accountability strategy for improving schools and raising achievement" (as quoted in Imig & Imig, 2008, p. 894). A flurry of lobbying ensued that would take the authority shift initiated by NCTAF in a new direction: "Instead of an emphasis on capacity building for education schools, Congressional efforts emphasized alternative certification and non-traditional routes into teaching. Professionalization gave way to accountability as the dominant policy frame for teaching and teacher education" (Imig & Imig, 2008, p. 894). The teacher education reform initiated by NCTAF had been swallowed by the momentum of the standards-based reform movement.

Accountability requires a metric, and the 1998 Reauthorization of the Higher Education Act required teacher preparation programs to report the passing rates of graduates on their states' teacher licensure assessments. This was the broader context in which the initial implementation of the Massachusetts MECT teacher licensure test occurred, as described in the Introduction to this study. Due to the inclusion of licensure testing requirements in its 1993 education reform law, Massachusetts was at the cutting edge of the wave of teacher licensure testing and standards-based reform. Similarly, when the NCLB Act of 2001 put subject matter knowledge at the heart of its definition of a highly qualified teacher, Massachusetts dramatically increased its requirements for subject matter coursework taken in arts and science departments (rather than education departments). A report by the Massachusetts Teacher Association's Center for Educational Policy and Practice noted major regulatory changes in 2001 and also traced a steady trend in Massachusetts licensure regulations over the last two decades that increased focus on content and decreased emphasis on pedagogy and practical experience (Center for Education Policy and Practice, 2008). The fact that teacher candidates with Bachelor's degrees can receive a preliminary license by simply passing the appropriate licensure exams cements the role of teacher knowledge as the sole gate-keeper to teaching in Massachusetts. For elementary and special education teachers, the new mathematics subtest has made that gate more difficult to get through than ever before.

Intent and content of the Massachusetts MTEL math subtest

The new licensure policies are designed to raise the standards of entry for elementary teachers in terms of their mathematics knowledge, including their pedagogical content knowledge. In April of 2007, the Board of Elementary and

Secondary Education (BESE) approved amendments to the regulations governing licensure so that they "focus[ed] on outcomes rather than on a list of arts and sciences coursework" (Chester, 2009, para. 1), according to Mitchell Chester, the Commissioner of MADESE. The new mathematics subject matter requirements for elementary teachers (regulation 603_CMR_7.06(7)(b)) read quite differently than the regulations for other subjects:

2. Mathematics

a. Basic principles and concepts important for teaching elementary-school mathematics in the following areas.

i. Number and operations (the foundation of areas ii-iv)

ii. Functions and algebra

iii. Geometry and measurement

iv. Statistics and probability

b. Candidates shall demonstrate that they possess both fundamental computation skills and comprehensive, in-depth understanding of K-8 mathematics. They must demonstrate not only that they know *how to do* elementary mathematics, but that they *understand* and can explain to students, in multiple ways, *why it makes sense.*

c. The Commissioner, in consultation with the Chancellor of Higher Education, shall issue guidelines for the scope and depth of knowledge expected in mathematics, described in a. and b. above.

Chester's memo from May 13, 2009, noted that these regulations represented "high expectations" (Chester, 2009b, para. 36) and represented an effort to "fundamentally

change the depth of mathematics competence among beginning teachers" (Chester, 2009b, para. 37).

The *Guidelines for the Preparation of Elementary Mathematics Teachers,* referenced in the regulation language, were released publicly in July, 2007, and approved by the BESE in December, 2007. The *Guidelines* were authored and championed by Tom Fortmann, a former engineer who had established himself as a mathematics-focused education reformer working as a full-time volunteer at Mass Insight Education, a prominent standards-based reform advocacy organization. Fortmann was appointed to the BESE in November, 2006, by the outgoing Republican Governor Mitt Romney. In an acknowledgement paragraph at the end of the cover memo of the *Guidelines*, the MADESE Commissioner at the time, David Driscoll, acknowledged Fortmann's work during the previous decade and also recognized the contributions of three mathematics professors. The policy changes were driven for and by mathematicians rather than education faculty, and the *Guidelines* explicitly address "mathematics department faculty" with an opening plea that is extraordinarily personal for a state agency guidance document:

> We--the candidates, the teacher preparation programs, and the Commonwealth-- need *you* to be teaching the courses referenced by these Guidelines. It is no longer someone else's problem that so many students and, ultimately, members of the workforce are ill-prepared for the challenges of an increasingly technological and competitive world economy. (Massachusetts Department of Education [MADESE], 2007, p. 2)

Between the disciplinary influence and the economic competitiveness rhetoric of statements like this, it would appear that the policies increasing the standards for teacher knowledge were in line with the deregulation movement and standards based reform.

Yet examination of the *Guidelines* reveal a push for capacity building with which advocates of professionalization would strongly agree. Although it called for increased contributions by mathematicians, it also encouraged stronger partnerships with education departments and recommended co-teaching models involving mathematics and education faculty. It called for preparation programs to be beefed up to three or four mathematics courses (versus the typical one or two) and cited a proposal from the National Council of Teachers of Mathematics, the professional mathematics teachers organization, to support the coursework recommendation. The *Guidelines* also heavily featured recommendations from the Conference Board of the Mathematical Sciences, an umbrella organization that brings together 17 other organizations expressly to promote cooperation between mathematicians and mathematics educators.

The pedagogical content knowledge (PCK) research of Deborah Ball, who is incidentally also a leader within the professionalization movement (see Ball & Forzani, 2009), was also strongly featured in the *Guidelines*. Her work was cited as a general reference, she was quoted directly about the importance and complexity of using student-friendly mathematical definitions, and her influence was clear from course design guidelines that explicitly drew attention to the importance of combining pedagogy and content:

> Pedagogy is typically the subject of "methods" courses, but separation of these two symbiotic topics is inefficient. Future teachers will be best served by math

professors who integrate mathematical principles, where appropriate, with

discussion of how these ideas can play out in classrooms, and by education

professors who ensure that methods are thoroughly grounded in content.

(MADESE, 2007, p. 9)

Test items from Ball's *Mathematical Knowledge for Teaching* instrument were reviewed

during the development of the new mathematics subtest, and the one open response item

on the practice test (worth 10% of the total test score) involved examining student work

and developing mathematical representations to promote student thinking, tasks right up

the pedagogical content knowledge alley. As the primary mechanism for implementing

the amended licensure regulations, the new mathematics subtest is intended to both raise

the standard of mathematical content, in line with deregulation advocates, and assess at

least some amount of pedagogical content knowledge, as promoted by professionalizers

who would emphasize the centrality of specialized knowledge for teaching.

Potential quality issues

Typically, the goal of teacher licensure and licensure exams are to ensure that

teacher candidates have a minimal level of basic competence (National Research Council

[NRC], 2001). The intention of the Massachusetts mathematics subtest seemed to go

beyond basic competence, however. Commissioner Chester's May, 2009, memo detailed

the process that was used to develop the new mathematics subtest and set the cut-score

(Chester, 2009b). The panel that helped set the cut-score were asked to consider the line

in the sand as "just acceptably qualified entry-level educators" (para. 32). Yet this panel,

as well as every panel at each stage of the process, was asked to attend closely to the

Guidelines document and the high expectations it embodied. This somewhat

contradictory message may have contributed to the initial pass rate of 27%. As for the pass rate, the Commissioner noted that "it was not a surprise to hear that many teacher candidates lack a strong math background" (MADESE, 2009, p. 6). Similarly, Fortmann found the "deficit in math knowledge among elementary teachers [to be] appalling" (MADESE, 2009, p. 6). The push for high standards and subsequent condemnation of teacher candidates for failing a basic test strongly echo the stance taken by state officials in the controversial aftermath of the original MECT teacher test implementation in 1998. In this case as well, although the emergency amendment allowed candidates to earn temporary licenses during a three-year transition period, the BESE did not back down from the testing requirements or official cut-score. On a broad level, this confusion about purpose threatens the face validity of the mathematics subtest.

In 1999, the United States Education Department commissioned the National Research Council to "examine the appropriateness and technical quality of teacher licensure tests currently in use and to consider alternatives for developing and assessing beginning teacher competence" (NRC, 2001, p. 2). Their *Testing Teaching Candidates: The Role of Licensure Tests in Improving Teacher Quality* report, released in 2001, noted licensure exam contractors typically focused on content validity (i.e., the exams test what they are intended to test). The report detailed a content validation process that matched quite well with the process laid out in Commissioner Chester's May 2009 memo, which involved stakeholders at each stage of test development, including writing the objectives, job analysis surveying of the field, item development and review, and setting the cut-score (NRC, 2001). The MADESE has apparently fulfilled the content validity requirements with a thorough and collaborative process.

It is less clear where the mathematics subtest stands in terms of empirically-based forms of validity. The *Testing Teacher Candidates* report recommended that tests should not be used to make licensure decisions before technical information on the performance of pilot tests has been analyzed and publicly released. At this time, more than four years after the first use of the test for licensure decisions, MADESE's testing contractor, National Evaluation Services (NES), has not released a technical report. Incidentally, the *Testing Teacher Candidates* report chided NES for not supplying technical information to its committee (the report did certify the technical soundness of the other major national vendor, Educational Testing Service). Until NES releases technical information, the public is left assuming that MADESE has verified the technical quality of the mathematics subtest.

One thing the MADESE has attended to in the past is the lower MTEL pass rates of non-white teacher candidates (see MADESE, 2007). The *Testing Teacher Candidates* report concluded that increasing expectations would likely decrease the diversity of the teaching force, and recommended that states keep this in mind when setting cut-scores (NRC 2001). Recent studies indicated that licensure testing does indeed decrease the number of non-whites in the teacher pool (Angrist & Guryan, 2008) and that non-white teacher candidates experience licensure testing as discriminatory (Bennett, et al., 2006). Some commentators have argued against licensure testing from the perspective that increasing non-white teachers is intrinsically valuable, particularly for urban schools (Johnson & Kardos, 2008; Villegas & Davis, 2008). Based on evidence of racial bias in other teacher licensure tests (Herbruck, 2006; Skiba, et al., 2008), and the uneven predictive validity found in Goldhaber and Hansen (2009),

MADESE should pay close attention to the impacts of the mathematics subtest on teacher candidates of varying demographics.

In their paper *Who is Teaching? Does it Matter?*, Zumwalt and Craig (2008) urged policymakers to assume a trade-off between teacher quality and diversity when setting requirements for licensure. Yet it is unclear whether licensure policies are related to teacher quality. Angrist and Guryan (2004) found that licensure testing policies did not improve the average SAT score of teachers, while Harrell (2009) showed that teachers whose transcripts suggested high levels of content knowledge in terms of course taking and GPA did no better on licensure exams in Texas than teachers with low levels of content knowledge. The fact that these studies show no association between licensure testing policies and traditional measures of teacher quality makes it difficult to properly consider the quality versus diversity trade-off.

Both the Zumwalt and Craig (2008) study and the *Testing Teaching Candidates* report recommended predictive validity studies that establish the relationship between teacher test scores and student achievement. In examining the predictive validity of the new mathematics subtest for elementary teachers, this study will contribute important information to the ongoing discussion of teacher testing in Massachusetts and adds to the knowledge base of teacher effectiveness research.

CHAPTER 3

METHODOLOGY

The two previous chapters have established the usefulness of studying the predictive validity of the mathematics MTEL subtest to shed light on the empirical validity of the test and inform related teacher quality policies and research. Predictive validity in terms of standards-based reform, which is the broader context in which the teacher test is based, means linking teacher test scores to student outcomes. This study will use approaches common to teacher effectiveness research, as defined by its conceptual framework and in line with the effects genre and outcomes-focused paradigm of teacher education research. Specifically, I will employ a quasi-experimental two-group comparison design within the context of a natural experiment. The analysis strategy will involve value-added growth models using cross-sectional data available from the state education agency. After describing the research design, I will discuss the analysis strategy in terms of data sources, Student Growth Percentiles as an outcome measure, and the specific regression models that will be employed. I will conclude by considering validity threats and ways to mitigate them.

Research Design

The overarching question of interest to this study is whether the Massachusetts MTEL teacher licensure math subtest is a valid predictor of teacher effectiveness. This study takes advantage of a natural experiment that has arisen due to a temporary situation

67

whereby teachers can receive licenses by passing the MTEL math subtest either fully or conditionally. Since June 2012, teachers have not received licenses based on conditionally passing the test, and therefore this study sample provides the chance to analyze the effectiveness of teachers who are currently ineligible for licensure. This allows us to look at the predictive validity of the cut-score in addition to providing a larger score range for examining the overall association of teacher test scores with student outcomes.

This study is a secondary analysis that uses pre-existing, cross-sectional, state-provided data. This is convenient and also sheds light on the validity of the math subtest and its related policies. By using data provided by the MADESE, I derive the control variables from information that the state already has on hand. This allows us to consider whether the math subtest represents useful information for licensure decisions above and beyond the information that the state has readily available. Although a comprehensive policy analysis is beyond the scope of this study, the use of state-provided data and outcome measures increases the relevance of results for policymakers.

Due to its quasi-experimental design, this study does not provide as strong of a platform for making causal inferences as a randomized experimental design. Randomized experimental designs are generally considered superior to other designs because the assignment to treatment conditions is exogenous - that is, the randomization makes it beyond the control of participants or researchers. This avoids selection bias and "renders members of the treatment and control groups equal in expectation prior to intervention" so that post-treatment between-group differences "must be a causal consequence of the intervention itself" (Murname & Willett, 2011, p. 136).

As a natural experiment that uses a two-group comparison design, this study will potentially provide the basis for making careful and qualified causal inferences. Murname and Willett (2011) specify three components to natural experiments:

> an underlying continuum along which participants are arrayed. We refer to this continuum as the "assignment" or "forcing" variable; an exogenously determined cut-point on the forcing variable that divides participants explicitly into groups that experience different treatments or conditions; and a clearly defined and well-measured outcome of interest. (p. 145)

This study exhibits each of these components. The assignment variable is teachers' scores on the MTEL math subtest. The cut-point is set by the state as a scale score of 240, dividing participants into full passing and conditional passing comparison groups in a way that is outside of their control as well as the researcher's. (Due to the temporary provision, there is a second cut point of 227 that divides teachers between conditionally passing and failing, but out study focuses on the main cut off due to sample size constraints.)

We can diagram this study in a way similar to Shadish, Cook, and Campbell's (2002) diagram of a Regression-Discontinuity design:

O_A C   X O_2

O_A C O_2

> where O_A is a preassignment measure of the assignment variable and C indicates that units are assigned to conditions on the basis of a cutoff score. That is, if j is a cutoff score on O_A, then any score greater than or equal to j entails being in one group, and anything less than entails being in the other. (p. 209)

In this study, condition X indicates those teachers who received a conditional pass and were temporarily allowed to teach before they fulfilled the permanent and official requirements of licensure. In a technical sense, the intervention that I am testing is the act of granting teachers a license who normally would not receive it. The other group fully passed and proceeded in a "business as usual" condition (i.e., without any intervention).

Natural experiments tend to be stronger designs than other observational studies. Shadish, Cook, and Campbell (2002) pointed out that only randomized and natural experiments provide a situation where the selection process is completely known, "so both designs can be viewed as special (successful) cases of selection bias modeling" (p. 224). The possibility of selection bias is greatly reduced in this study because participants are assigned to groups directly based on whether their test score lies above or below the cutoff of 240. As an assignment variable, any error that exists in the math MTEL test as an instrument to measure an underlying trait (e.g., content knowledge) is inconsequential. As an assignment variable, the MTEL score does not represent an underlying trait - each teacher's score either lies above or below the cutoff and assigns those teachers to a group accordingly. From the policy's standpoint, the groups are not equal in expectation because the cut-score is assumed to be meaningful in terms of teacher effectiveness. That is, the full pass group is expected to be more effective than the conditional pass group. For the purposes of this study, I define the null hypothesis as assuming that the two groups are the same in terms of effectiveness and therefore that there is no association between teacher test performance and student performance.

Analysis Strategy

As detailed in the literature review in Chapter 2, value-added growth models are well established as the preferred approach in teacher effectiveness research. The most typical strategy, taken from econometric methods, involves the use of educational production functions that include a measure of student outcomes on one side of the equation and various inputs on the other side of the equation. The inclusion of teacher variables provides the "value-added" aspect of this study, which includes teacher math MTEL test performance as the key independent variable along with other teacher and student characteristics as controls in order to isolate the impact of the teacher licensure test. This study qualifies as teacher effectiveness research (as defined by the conceptual framework) by using a measure of student growth as the dependent variable. After summarizing the data sources and variables, discussing the outcome measure, and detailing the process by which the study's data file was prepared, the section will end by specifying the regression models.

Data Sources

The MADESE provided all the data for this study. A data request was submitted to the Office of Strategic Planning, Research, and Evaluation in February, 2011 and an initial Memorandum of Understanding was established in spring 2011 for the pilot study that was conducted in the summer of 2011. An addendum to the MOU was established in November, 2011, for up-to-date data to be provided in January, 2012. The sources are listed below along with the definitions of each variable that will be used in the statistical models.

(1) Education Personnel Information Management System (EPIMS), SY2010-2011

data for teachers in grades 4 and 5. This provides teacher background characteristics. I have also accessed data from the previous two years in order to look at experience.

(2) Student Course Schedule (SCS) pilot data, SY2010-2011 data for students in grades 4 and 5. This allows teachers to be linked to the students in their classrooms.

(3) Student Information Management System (SIMS), SY2010-2011 data for students in grades 4 and 5. This database includes background characteristics on students.

(4) 2011 MCAS data for students in grades 4 and 5. This provides math Student Growth Percentile (SGP) scores and MCAS scale scores.

(5) Educator Licensure and Recruitment (ELAR) data for the teachers in EPIMS. This provides teacher license test results, including MTEL scale scores, as well as other background characteristics not included in EPIMS such as all licenses held and detailed information about college degrees.

The data therefore includes extensive current and historical information on students and teachers. Table 1 summarizes the variables used in regression analyses in this study.

Data File Preparation

The data sources were merged following a process recommended by the MADESE Office of Research and Evaluation. After describing how the data sources were compiled, I detail how certain variables were created. Descriptive statistics for the teachers and students included in the final data file will be presented in Chapter 4.

Data compilation process. The initial SCS file included 7,747,838 records. These records represented all courses for Massachussetts students during the 2010-11

Table 1

Summary of Variables Included in Regression Analyses

Variable name	Type	Possible values (Notes)	Data source
Student variables			
Math SGP[a]	Continuous	1-99	MCAS
African-American	Categorical	Dummy variable	SIMS
Hispanic	Categorical	Dummy variable	SIMS
Low income	Categorical	Dummy variable (free and reduced lunch)	SIMS
Limited English Proficient (LEP)	Categorical	Dummy variable	SIMS
Special education (SPED)	Categorical	Dummy variable	SIMS
MCAS score[b]	Continuous	200-280 (2011 math MCAS scale score)	MCAS
Prior achievement	Continuous	200-280 (2010 math MCAS scale score)	MCAS
Student classroom variables			
Class size	Continuous	1-30	SCS
Peer effects	Continuous	200-280 (average 2010 math MCAS scale scores of students in the classroom)	MCAS
Teacher variables			
Median math SGP[c]	Continuous	1-99	MCAS
MTEL math test score	Continuous	200-280	ELAR
MTEL math subtest pass status	Categorical	Dummy variable (pass, conditional pass, fail, or pass, not pass in some analyses)	ELAR
Experience greater than 2 years	Categorical	Dummy variable (two years or less, greater than two years)	EPIMS
Possesses graduate degree	Categorical	Dummy variable	ELAR
Holds preliminary license only	Continuous	Dummy variable	ELAR
Teacher classroom variables			
Percent of students SPED	Continuous	0-100	SCS
Percent of students low income	Continuous	0-100	SCS
Percent of students LEP	Continuous	0-100	SCS
Prior achievement	Continuous	200-280 (average 2010 math MCAS score of 2011 students)	MCAS

Note. SGP = Student Growth Percentile. MCAS = Massachusetts Comprehensive Assessment System. SIMS = Student Information Management System. SCS = Student Course Schedule. MTEL = Massachusetts Test for Educator Licensure. ELAR = Educator Licensure and Recruitment. EPIMS = Educator Personnel Information Management System.
[a] Independent variable in main student-level regression analyses. [b] Independent variable in supplemental student-level regression analyses. [c] Independent variable in main teacher-level regression analyses.

school year. This was the first year that SCS data was collected statewide by the

MADESE. The SCS file was trimmed to 843,649 records by selecting only those records

that included courses in which mathematics was taught (i.e., the "course" variables were

coded to values that correspond to either "elementary multi-subject," "mathematics

middle school," "foundation mathematics middle school," "pure mathematics middle

school," or "Algebra I grade 8"). Each of these course records represented one student

and contained unique identifiers that allowed those students to be linked to their teachers.

Next, the two student data files were prepared and then merged into the SCS

file. The MADESE provided a file with student MCAS test data for all students in grade

four and five during the 2010-11 school year. Using the State Assigned Student

Identifier (SASID) as the linking variable, the MCAS file was merged into the SCS file.

The 144,365 records in the MCAS file, each representing one student, populated 196,338

records in the SCS file, because some students showed up in multiple mathematics

courses (e.g., if a student was taking support or enrichment mathematics courses in

addition to their main mathematics course). This 196,338 was about 23% of the original

843,649 student course records in which mathematics was possibly taught in elementary

and middle school, which was reasonable because the MCAS data included only grade

four and five students. I finished the student base file by merging in the other student

data file, the Student Information Management System (SIMS) file, which contained the

official student demographic data. Finally, I created a unique course code variable by

concatenating four variables (course location, course code, section code, and term) that

provided a unique identifier for merging this student base file to the teacher base file.

To prepare the teacher base file, I started with the Educator Licensure and

Recruitment (ELAR) file. As provided by MADESE, the ELAR file initially represented just grade four and five teachers during the 2010-11 school year, organized so that each record represented one instance of a teacher having taken an MTEL licensure test. By selecting only those records that included General Curriculum Mathematics (i.e., the "MTEL mathematics subtest") data, the 31,204 total records were trimmed to 684 records. Since in some cases a teacher had taken the MTEL mathematics subtest multiple times, the duplicates were deleted by selecting the 407 unique teacher records that included the highest score on the MTEL mathematics subtest. This means that out of approximately 9,000 educators who were working with grade four and/or grade five students during the 2010-11 school year, 407 of them, or less than 5%, had taken the MTEL mathematics subtest since it had started in March, 2008.

Once the teachers who had taken the MTEL mathematics subtest had been identified, I merged in the Educator Personnel Information Management System (EPIMS) file that contained demographic and other data about these teachers. The EPIMS file was cumulative, so that each record represented a unique teaching assignment during the 2010-11 school year; a single teacher would have multiple records if he or she had taught more than one course (note that this could be in different schools) during the school year. The EPIMS file provided by the ESE contained 18,040 course records for the approximately 9000 educators who had taught grade four and/or grade five students during the 2010-11 school year. Once the files were merged, the 407 teachers from the ELAR file were represented by 663 course records because some teachers taught multiple courses. A unique course code was created by concatenating the same four variables as the student base file so that the 663 courses in the teacher base file could be

merged into the student base file.

When the 663 course records from the teacher base file were merged into the 196,338 course records from the student base file, there were 2897 matched course records for which there was full data on a unique student and teacher combination. Among the 2897 records, each corresponding to a unique student, there were 130 unique teachers linked to 166 unique courses (because some of the teachers taught more than one course, probably in a middle school model). Of the 2897 students, 257 students were dropped because they did not have a math SGP score. The remaining 130 teachers and their 2640 students in the final data file represented the study subjects; descriptive statistics are provided in Chapter 4.

It is worth considering why the 407 teachers from the base teacher file decreased to 130 teachers in the final data file. The reporting rules for the SCS data collection required that teachers were classified as either "teachers" or "co-teachers". Of the 407 teachers, 204 teachers were classified as "co-teachers" and hence dropped out of the study because they were not involved in a unique student and teacher combination; it would be impossible to attribute the impact of these teachers on their students' scores. The remaining 203 teachers were classified as "teachers" that theoretically should have been linked to unique student data. In discussions with the MADESE Office of Research and Evaluation, it became clear that the validation rules were not firm during this first year of statewide SCS collection, and therefore it is reasonable that only 128 of these 203 teachers were associated with unique student and teacher combinations. The fact that of the 203 teachers, 135 were assigned as classroom teachers and 68 were special education teachers, with the final 130 split as 113 classroom teachers to 17 special education

teachers, does suggest that the majority of the lost teacher participants were in roles as special educators where they were less likely to have been teaching students solely. As the SCS data collection and data cleaning processes become more robust in the future, it should become more clear which teachers are responsible for which students. For the purposes of this study, I have limited the sample to unique combinations of teachers and students so that I can more directly attribute teacher impacts on student scores.

Creation of variables. Many variables were in a useable format straight from the files provided by the MADESE. Some variables required simple transformations to create dummy variables that could be used in the regression models, such as student race and teacher degree. Yet other variables were created using straightforward calculations, such as the number of students in each course, the percentage of students from particular student groups assigned to each teacher, or the median SGP for each teacher.

The creation of the teacher experience variable, on the other hand, was a bit more complicated due to the fact that the MADESE does not track when teachers initially started teaching in Massachusetts. The EPIMS file provided the date that each teacher was hired by his or her current district. In some cases the hire date was an accurate marker of teacher experience. If a teacher had not taught in the same district for their entire career, however, the teacher experience variable underestimated the number of years of experience. MADESE has collected teacher data since the 2007-2008 school year, allowing teachers who had 1-3 years of experience to be verified. For any teachers with a hire date during the 2008-2009 school year, 2009-2010 school year, or 2010-2011 school year, which would suggest between 1-3 years of experience, the statewide database was consulted to see whether these teachers showed up in a different district

during a prior year. If a teacher did not teach in a different district during the school year before their listed hire date, it was assumed that the teacher had started their teaching career at the hire date. For teachers with a hire date during the 2007-2008 school year or prior, no confirmation was possible.

The teacher experience variable therefore has two noteworthy problems. First, it is possible that a teacher could have re-entered the Massachusetts work force and been erroneously assigned a value of 1-3 years of teaching experience. Secondly, and more importantly, many teachers who are assigned a value equal to or greater than four years of teaching experience have probably actually taught more years than reflected by that value, since the value is calculated based on the hire date in their current district. These problems are slightly mitigated by the fact that the literature suggests that teacher experience only matters within the first 3-5 years. Assuming that the benefits of experience may plateau after two years, and because the teacher experience variable is most accurate within the first three years, the teacher experience variable is operationalized as a categorical variable (two or fewer years versus more than two years experience).

Regression Models

This section discusses the independent variable, the independent variables, and specifies the variables that are used in each of four main regression models.

The advantages of SGP as the dependent variable. Starting in 2009, the MA MADESE has published an SGP statistic for each student. SGP utilizes quantile regression methods (see Betebenner, 2009; Betebenner & Linn, 2010) to measure student progress by comparing changes in a student's MCAS score to changes in the MCAS

scores of his or her academic peers. Each student's MCAS score is placed on a

distribution of students with similar achievement profiles (i.e., prior MCAS scores),

allowing the student's MCAS score to be assigned a percentile ranking from 1 to 99.

In an informational publication on SGP, MADESE described the rationale for

SGP by appealing to fairness:

> Each student is being compared to his or her academic peers: other students
>
> statewide with similar MCAS test score histories. This makes for a fair
>
> comparison because it allows us to describe the likely range of scores observed
>
> among all students with a similar MCAS test score history, and therefore to see
>
> how quickly the student improved given his or her past test scores. (MADESE,
>
> 2011, p. 17)

SGP scores are increasingly relied upon by the MADESE for high-stakes purposes.

Median SGP scores have been used for school and district accountability starting with

data from school year 2010-11 to make determinations about accountability status.

Median SGP at the teacher level is required to be incorporated into Educator Evaluation

systems by school year 2013-15. Clearly, MADESE is proceeding with the assumption

that SGP as a growth measure provides a fair representation of student performance and

that it is stable and robust enough to use for high-stakes decisions.

Many value-added growth models, such as the William Sander model in

Tennessee, require tests to be vertically aligned. By using the normative approach of

percentiles, SGP avoids this requirement and therefore provides a growth statistic for a

state like Massachusetts that does not have vertically aligned assessments. While a

model like Sander's looks at how student performance diverges from a prediction, SGP

simply reports how students actually perform compared to their academic peers. Betebenner (2009) argued that this is advantageous from an accountability standpoint because it avoids the complexities of statistical prediction and is more intuitively understandable to the public due to the widespread use of percentiles in pediatrics.

As a measure of student learning growth, SGP potentially possesses many beneficial characteristics. Because SGP incorporates the past achievement of students, it should represent accumulated achievement factors that led up to the original test score in grade 3, such as early childhood experiences. MADESE suggested that student background may not be correlated with growth:

> Research shows that there are correlations between a student's demographic group and their performance on the MCAS. Is the same true with growth? Not necessarily. The relationship between demographics and growth is complex, much more so than the relationship between demographics and achievement. For instance, because there are numerous studies that have established a correlation between economic disadvantage and achievement level, one might expect that low-income students would achieve at a lower *level* than students without such economic disadvantages. However, it is not so clear that low income students should *grow* slower once you've taken performance level into account, given the way we calculate growth. (MADESE, 2011, p. 19)

Furthermore, because each student's growth is a normed comparison to their academic peers, SGP should account for other unmeasured factors that may be similar across the academic peer group (e.g., inherent motivation). And since the academic peer group is redefined each year based on the additional information of the prior year's score, a

80

student's SGP should be fairly independent from year to year. In theory, SGP is a very useful outcome measure for measuring student growth in a research context.

Yet, although several states are now reporting SGPs associated with their state-wide assessments, I was not able to find any observational studies that use SGP as an outcome measure. Wright (2010) simulated several value-added models with the same data set and confirmed that SGP was robust to outliers and, like other growth measures, not nearly as strongly correlated to student background as status measures of achievement. Although Wright found other measures to be superior in terms of stability, SGP was found to be a fairly typical growth measure that would be expected to perform well in a research context. In a descriptive study, Slaughter (2008) chose SGP as the best measure to describe middle school achievement growth because it allowed examination of growth across the entire spectrum from very low to very high achieving students. Overall, however, the use of SGP for research purposes has been very limited, probably due to its fairly recent dissemination (D. Betebenner, personal communication, November 23, 2011). Beyond it's strong potential as an outcome measure, the MADESE's reliance on SGP makes it particularly appropriate for use in studying MADESE policy.

Independent variables. Due to SGP's relatively weak correlation with student background and its use by the state in educator evaluation, I include teacher-level models that are relatively parsimonious. These models use an average (median) SGP score for each teacher as the outcome measure and include independent variables at the teacher level only, which makes the models easily interpretable but also less powerful because they do not make use of all of the available data. The student-level models predict

student math SGP and include both teacher and student background control variables, chosen due to support in the literature (e.g., teacher experience) or common practice (e.g., SES, class size).

Both the teacher-level and student-level models are analyzed using two different versions of the key independent variable, teacher test score: MTEL math subtest pass status and MTEL math subtest score. The combinations of the two versions of the key independent variable (MTEL math pass status and score) and the two levels (teacher and student) produce the four main regression models described below.

Model 1a. Predicting teacher median SGP by teacher MTEL math subtest pass status. The regression model can be written as:

Teacher median SGP = classroom variables + teacher variables

Where classroom variables include:

- Percentage of students with disabilities

- Percentage of students who are low income

- Percentage of students who are Limited English Proficient

- Prior achievement of teacher's students

And teacher variables include:

- Teacher experience greater than two years (dummy variable)

- Teacher holds preliminary license only (dummy variable)

- Teacher possesses a graduate degree (dummy variable)

- Teacher MTEL math test fully pass (dummy variable)

This model looks at the association of teacher outcomes on MTEL and the median SGP of their students. It sheds light on whether teachers in the full pass group have higher

average student growth than teachers in the conditional pass group, controlling for other teacher and classroom characteristics. By not including student background characteristics, this model mirrors the way that median SGP will be used for teachers in the state's Educator Evaluation system.

Two variables were discarded based on results from a pilot study conducted during the summer of 2011. The first was "teacher mathematics degree", which looked at whether a teacher had majored in mathematics during postsecondary schooling. Out of the 130 teachers in the study sample, only two had a degree related to mathematics, so this variable was considered superfluous. The second was teacher results on the MTEL Communications and Literacy licensure test, which was originally conceived as a possible proxy for cognitive ability. All 130 teachers had passed this test and so there was not enough variation in the measure to serve as a useful covariate. Regression analyses including these variables can be found in Appendix A.

Model 1b. Predicting teacher median SGP by teacher MTEL math subtest score. This regression model is the same as 1a except that it uses teacher score on the MTEL math test instead of pass status as the key independent variable.

Model 2a. Predicting student SGP by teacher MTEL math subtest pass status. The regression model can be written as:

Student math SGP = student background variables + classroom variables + teacher variables

Where student background variables include:

- African-American (dummy variable)

- Hispanic (dummy variable)

- Male (dummy variable)

- Low income (dummy variable)

- Limited English Proficient (dummy variable)

- Special education (dummy variable)

Classroom variables include:

- Class size

- Peer effects, prior achievement of students in the class

And teacher variables are the same as those used in the teacher-level models:

- Teacher experience greater than two years (dummy variable)

- Teacher holds preliminary license only (dummy variable)

- Teacher possesses a graduate degree (dummy variable)

- Teacher MTEL math test fully pass (dummy variable)

This model looks at whether a student's growth in mathematics is associated with his or her teacher's pass status on the MTEL math test. The model is analyzed in four blocks, with block one focused on student background variables only, block two adding classroom variables, block three adding teacher characteristics, and block four representing the complete model that includes teacher licensure test performance. This allows us to analyze the impact of each of these different categories of variables and ultimately gain insight into what the teacher licensure test results provide in terms of additional information.

Model 2b. Predicting student SGP by teacher MTEL math subtest score.
This regression model is the same as 2a except that it uses teacher score on the MTEL math test instead of pass status as the key independent variable. Similarly, it will be

implemented in four blocks of variables.

Validity Threats

Shadish et al. (2002) defined validity as "the truth of, correctness of, or degree of support for an inference" (p. 513). In the case of this study, I am investigating inferences at two levels. From a policy standpoint, I am trying to support an inference about the predictive validity of the MTEL math subtest by looking at whether there is an association between teachers' performance on the MTEL math subtest and the outcome of their students on MCAS. In this case I rely on the state's built-in assumptions about its own measures. From a research standpoint the inference I am attempting to support is more demanding. Put in terms of the conceptual framework, this inference pertains to teacher effectiveness based on the assumption that the MTEL math subtest measures teacher content knowledge and the assumption that the MCAS measures student learning. From this perspective, the inference is about a causal relationship between teacher knowledge and student learning. Shadish et al. cited John Stuart Mill's logic of causal relationships: "a causal relationship exists if (1) the cause preceded the effect, (2) the cause was related to the effect, and (3) we can find no plausible alternative explanation for the effect other than the cause" (p. 6). This study clearly satisfies the first condition. Since I am using a quasi-experimental design rather than a randomized experimental design, I need to carefully consider validity threats to address the second and third conditions. Following Shadish et al., I will consider four types of validity threats in turn: statistical conclusion validity, internal validity, construct validity, and external validity. This section will be most relevant to the causal inference that I am trying to support

within the research context, though all considerations will apply to the less demanding policy context inference as well.

Statistical Conclusion Validity

Statistical conclusion validity concerns inferences about the existence and magnitude of covariation between two variables. In this study the variables of interest are teachers' MTEL math subtest performance and their students' outcomes on MCAS. I will test the null hypothesis that these two variables do not covary (i.e., that they are not associated) using the multivariate regressions detailed in the "Analysis Strategy" section of this chapter. If I reject the null hypothesis, I will seek to estimate the magnitude of the association between these variables.

The first relevant threat of this type is *low statistical power.* A study must be sufficiently powered to detect covariation and therefore avoid a Type II error (incorrectly concluding that there is no association when there actually is). The key consideration for statistical power is the sample size (Murname & Willett, 2011). The sample size of this study is 130 teachers and 2640 students.

In order to assess the validity threat posed by an underpowered study, I conducted a basic *a priori* statistical power analysis using the "linear multiple regression: fixed model, single coefficient" test within the G*Power 3.1 power analysis program (Faul, Erdfelder, Lang, & Buchner, 2007). In addition to sample size, statistical power is determined by two other factors. One is the alpha level, which will be set at the conventional level of .05. The other is the effect size. The literature suggests that the effect size of elementary mathematics teacher licensure test outcomes range from a small effect size of .015-.047 for the score itself (Clotfelter, Ladd, & Vigdor, 2010; Goldhaber

& Hansen, 2009) to a medium effect size of .05 for pass status (Goldhaber, 2007; Goldhaber & Hansen, 2009). For the small effect size, the sample size required for adequate statistical power is 869, while the larger effect size requires a sample size of 29. Therefore I can be confident that the study is sufficiently powered for all student-level models (2a and 2b) but it may be underpowered to detect a small effect size at the teacher level (models 1a and 1b).

Shadish et al (2002) pointed to other important factors for statistical power that are not included in the power analysis. One factor that decreased the power of this study is the fact that the comparison groups were unequal in size. The inclusion of covariates in the statistical models, however, increased power. The fact that the measures are standardized and used to make high-stakes decisions means that they have high reliability, which is also beneficial for statistical power. Overall, the relatively large sample size provides the best defense against the threat of an underpowered study.

The use of powerful statistical methods is a benefit insofar as I can ensure that their assumptions are met and avoid the threat of *violated assumptions of test statistics*. One assumption that I need to address is normality. At the student level, the outcome variable of SGP was uniformly distributed, so I converted SGPs into Normal Curve Equivalents (NCEs) to make them into scale scores that were normally distributed. I also need to be mindful of the key violation that Shadish et al (2002) highlighted, unit of analysis, which is a common dilemma in education research because of the fact that different units of analysis are naturally nested within each other. Nesting can make it difficult to tease out effects of variables at one level (e.g., teachers) on units at a lower level (e.g., students) because the lower level units are similar in ways that have nothing to

do with the higher level variables (e.g., high-achieving students may cluster together in a classroom due to factors that have nothing to do with the teacher).

Traditionally there have been two main analytical options for dealing with nested data. One option is to aggregate all the data to the higher level, but this does not utilize all available data and misses the within-group variation. This option increases the possibility of Type II error, in which an effect goes undetected. The other option is to disaggregate all data to the individual level, but this violates the assumption of independent observations (for example, all students in the same classroom would have the same teacher), and results in an increased likelihood of Type I error, in which an effect is detected that does not actually exist. Although these approaches may be somewhat statistically defensible and will yield results, the basic problem lies with the inferences that can correctly be drawn from these results when "relationships discovered at one level are inappropriately assumed to occur in the same fashion at some other (higher or lower) level" (Luke, 2004, p. 6). A third option, more commonly used recently, is the use of multilevel modeling techniques that are statistically sophisticated but somewhat difficult to interpret. In this study I will run models at the aggregated (teacher) level and the disaggregated (student) level, keeping in mind the biases of each in the interpretation of results.

Somewhat related to the issue of nesting is the strong possibility that nonrandom assignment has distributed *unmeasured variables* among the study participants in a way that varies systematically alongside the variables of interest. This could confound the results. If, for example, teachers who fully passed the math MTEL subtest tended to be assigned students who were more internally motivated than teachers who conditionally

passed, I may erroneously conclude that there is an association when there is none (i.e., a Type I error). This study has a good starting point because it includes variables that represent most of the major factors identified in teacher effectiveness research (as discussed in Chapter 2): teacher experience, license, degree, and major. The student-level models also include background characteristics of teachers and students. In addition to being convenient and policy relevant to focus on variables for which data is available from the state, it is also defensible from a validity standpoint because all of the key covariates will be present in the models.

Recall from Chapter 2 that many recent studies of teacher effectiveness employ fixed effects to deal with the problem of nonrandom assignment and the associated validity threat of unmeasured variables (for example Betts, et al., 2003; Clotfelter, Ladd, & Vigdor, 2007; Goldhaber & Hansen, 2009). Fixed effects models typically use multiple measurements over time to use students as their own control by focusing on variations within individuals rather than between individuals or groups. This requires variation of the variables of interest within participants over time (Allison, 2009), which was not available in the data set. Since the policy came into effect only recently, there would be very few students who had a teacher who had fully passed in one year and then a teacher who had conditionally passed in the next year.

The nature of the key variables should help mitigate against the possibility that students were assigned to teachers in a way that would produce an unwarranted effect. On the input side, MTEL pass status is much more subtle than a variable such as experience or even college major. Whether a teacher fully or conditionally passed MTEL is unlikely to be known by hiring administrators (i.e., they will only know that the teacher

89

is licensed), much less factored into decisions about student assignment. Using SGP as the outcome measure should also help, since the normed comparison of growth to academic peers should mean that students' growth is sensitive to their learning in a given year. Students themselves are not inherently "high growth" or "low growth" and so their assignment to teachers should matter less than with other outcome measures. Therefore, although I am left to deal with nonrandom assignment by controlling for the key covariates, the issue of unmeasured variables may be less damaging than in other teacher effectiveness studies.

Internal Validity

Internal validity refers to "the validity of inferences about whether the relationship between two variables is causal" (Shadish, et al., 2002, p. 508). Internal validity threats usually take the form of alternative causes to the one of interest. In the case of this study, if I find that teacher MTEL math test scores are statistically associated with student growth scores, supporting a causal inference linking teacher knowledge to student learning, I increase internal validity to the extent that I can rule out other causes that could give rise to this statistical association.

The most prominent internal validity threat in quasi-experimental studies is *selection,* in which the comparison groups differ systematically in a way that confounds the association of interest. Selection is not an issue when random assignment is used because randomly formed groups differ only by chance and so it is extremely unlikely that those differences would coincidentally be systematic in a way that manifests as the association I am examining. Although "selection is presumed to be pervasive is quasi-experiments" (Shadish, et al., 2002, p. 56), this study has the advantage of being a natural

experiment so that selection actually becomes a strength. As noted previously in the "Study Design" section of this chapter, the fact that teachers were assigned into one of the comparison groups based on their MTEL pass status makes this a very special selection process due to its full transparency and lack of ambiguity. Many of the selection issues that plague other quasi-experiments, such as the potential for participants to manipulate their way into (or out of) treatment, or designs in which participants volunteer into treatment, introduce endogenous selection issues that raise the possibility that comparison groups differ systematically from the outset in a way that biases the results. The use of a clear externally determined cut-score to assign participants to comparison groups means that this study takes advantage of an exogenous selection method that is out of the hands of participants and researchers and can't be manipulated in unknown ways. As long as the MTEL math test was scored and recorded properly, the two comparison groups should be perfectly formed in relation to their scores. The reliability of the test to accurately tap into an underlying trait, and variability issues due to measurement error, are important in terms of construct validity but are not relevant to selection per se. Similarly, selection issues can often give rise to unmeasured variable threats, which were discussed in the previous section. The status of this study as a natural experiment mitigates against the strict selection threats that hamper many quasi-experimental studies.

Once the initial selection has been made, however, *attrition* threats may arise. The population of interest is comprised of all candidates who passed the MTEL math subtest. Based on cut-score, some of them were offered a teaching license and others were offered a conditional license. A subset of the test-passers ended up in classrooms

and had student results that allowed them to be included in the study sample. The attrition threat biases the results of this study to the extent that attrition from the group that received the "treatment" of being offered a conditional license is systematically different (in a way that parallels the association of interest) than the attrition from the group that was offered a full license. It is possible, for example, that some participants who were offered the conditional license were so discouraged at the prospect of having to fully pass the test within three years that they decided to avoid the teaching profession. In this case the ultimate subset of the conditional group would probably end up stronger than its initial composition, thus attenuating effects. One could also imagine scenarios that could weaken the fully passed group, such as the attrition of participants who have better alternatives elsewhere, that could artificially inflate the effect. Therefore it is important that this study look at the likelihood of an attrition threat by closely examining the data of the study sample as compared to other teachers; this could provide evidence of attrition and whether the subset of participants for which I have student outcomes are demonstrably different than the full pool of participants.

This study avoids many of the other possible internal validity threats. *Testing* can be an issue in pre- and post-testing designs, but does not come into play in this study. Teachers can take the MTEL math subtest multiple times, and it is possible that exposure to the test could help them surpass the cut-score and jump into the fully pass group. Perhaps most importantly, the threat of *ambiguous temporal precedence* is not an issue. In all cases, teachers will have taken the MTEL math subtest before they taught the students who took the MCAS to produce the outcome measure. Although a randomized

experiment would clearly be preferable for internal validity, overall this study provides a solid platform for making causal inferences relative to other quasi-experiments.

Construct Validity

Construct validity "involves making inferences from the sampling particulars of the study to the higher-order constructs they represent" (Shadish, et al., 2002, p. 65). Construct validity does not strongly apply to the inferences that this study will make at the level of policy. That is, on a superficial level, from the perspective of policy I am simply looking at the association of MTEL math subtest scores with student SGP on MCAS. I don't have to justify these measures to policymakers because they are state-determined and so their construct validity is accepted by the state already. From a research perspective, however, for the results of this study to be generalizable, the constructs must be clearly defined and assessable by the measures used. The rest of this section will assume that construct validity is high from a policy perspective and focus on construct validity threats that are relevant to inferences in the research context.

First, it is important to clearly define the constructs at play so that I can address the threat of *inadequate explication of constructs*. This study focuses on two categories within the conceptual framework that are both under the heading of "inputs": teacher qualifications and teacher characteristics. These inputs are the things that a teacher brings to the classroom as internal resources. The predictor variables represent the inputs that I have operationalized in order to include in statistical models. Some of the teacher characteristic inputs are fairly straightforward in terms of the constructs that they represent, such as teacher gender and race. The variable of "teacher score on MTEL Communication and Literacy test", however, deserves explication. This variable is meant

to represent the construct of cognitive ability, which has been identified in the literature as a possibly important covariate. Some studies have used tests such as the SAT as a proxy for cognitive ability (Boyd, Lankford, Loeb, Rockoff, & Wyckoff, 2008), and some have used less direct proxies such as college selectivity (Wayne & Youngs, 2003), but more commonly cognitive ability is excluded from teacher effectiveness models. The MTEL Communication and Literacy test is a good proxy for cognitive ability in this study because it should test different skills than the MTEL math subtest. It is also the one test that is required for all educator licenses, so its inclusion is helpful in parsing out the information (effect) contributed by the MTEL math subtest. In order to support its usefulness as a proxy measure of cognitive ability, this study will examine the teacher participants' MTEL Communication and Literacy Test to rule out ceiling effects and strong correlation to the MTEL math subtest.

The other category of inputs, teacher qualifications, contains the key predictor variables of teacher MTEL math subtest score, degree, license, and experience. Experience within the first 3-5 years is strongly supported in the literature as associated with teacher effectiveness. This study operationalizes experience by looking at a database to see whether teachers were teaching in Massachusetts up to three years ago, so this variable is defined as whether teachers have two years or less teaching experience versus more than two years. Degree and license are inconsistently supported in the literature, but they are common paper qualifications used in teacher effectiveness research and for hiring decisions. The possession of a mathematics degree would have been included as a covariate but it is not present enough in the sample for inclusion (only 2 of 130 teachers had a mathematics degree). As for the math MTEL test, the fact that

94

the process for developing the test was robust and transparent suggests content validity that provides the benefit of the doubt that it is measuring its intended constructs, content knowledge and pedagogical content knowledge.

Once the constructs have been explicated and tied to the operationalizations in this study, the second threat to construct validity that must be considered is *mono-operation bias*. This threat arises due to the fact that "any one operationalization of a construct both underrepresents the construct of interest and measures irrelevant constructs, complicating inference" (Shadish, et al., 2002, p. 73). Although the MTEL math subtest has a certain amount of content validity, no technical reports are available to confirm that it exhibits the basic characteristics of a high quality assessment (e.g., reliability). If this study finds support for the null hypothesis, suggesting a lack of predictive validity of the MTEL math subtest, it will be unclear whether the test is simply a poor measure of the construct of math content knowledge or if there is actually no association between the construct and student outcomes in this instance. Without a second measure (operationalization) of the construct of teacher content knowledge, I have no way to triangulate the inclusion of the construct as an input.

This study also has an over-reliance on a single assessment, MCAS, for the student outcome measure. This problem has plagued teacher effectiveness research. Koretz (2008) argued that growth models based on annual standardized tests are problematic due to timing – a teacher or school only impacts a student from September until the annual testing period (May for the math MCAS), whereas some amount of unattributable growth may occur (or doesn't, as the case may be) between May and September. He made similar points about other commonly cited problems with

95

standardized tests, including sampling errors (i.e., the test cannot reflect with absolute accuracy the knowledge of one or more students, and this is a bigger problem for smaller groups of students); narrowing the curriculum (no single test can measure the full domain of important knowledge, and educators are more likely to emphasize what *is* measured); and coaching (teachers may spend time on training students in test-taking skills). Although MCAS is widely considered a high-quality assessment, as a standardized test it brings certain challenges and the results of this study must be interpreted in the context of these limitations.

The last construct validity threat to consider is *confounding constructs with levels of constructs*, which happens if a researcher "draw[s] a general conclusion about constructs that fails to recognize that only some levels of each facet of that construct were actually studied and that the results might have been different if different levels were studied" (Shadish, et al., 2002, p. 76). Recall that the panel that helped set the cut-score on the MTEL math subtest were asked to consider the line in the sand as "just acceptably qualified entry-level educators" (Chester, 2009b, para. 32). It is reasonable to assume that the participants who didn't reach that cut-score have moderate to low math content knowledge. The conditional pass group widens the net to include teachers with lower scores than are allowed under the official cut-score; this increases the teacher pool only for low scorers. Furthermore, the test was designed to make determinations most accurately around the cut-score, so it is likely to be less accurate and meaningful for scores far above the cut-score. Therefore when interpreting results, it is important to acknowledge that the construct of teacher mathematics content knowledge is not

represented across a full spectrum from very low to very high, but likely only includes levels that are relatively low to moderate.

External Validity

External validity "concerns inferences about the extent to which a causal relationship holds over variations in persons, setting, treatment, and outcomes" (Shadish, et al., 2002, p. 83). The common formulation is that internal validity is about whether a causal inference is sound and external validity speaks to the extent to which it can be generalized. Whereas the internal validity of this study is high from a policy perspective, the external validity is rather low when I think of this study as focusing on a particular teacher licensure policy that uses a state-specific teacher licensure test and examines the association with a state-specific student assessment. Although the causal inferences of this study are stronger from a research perspective, and therefore more demanding in terms of the validity threats discussed so far, causal inferences within the research framework will tend to be more generalizable and thus have higher external validity than causal inferences within the policy framework. Even so, it is important to consider external validity threats in order to explicate the limited generalizability of this study.

Causal inferences have limited generalizability due to *interaction of the causal relationship with units*. This study focuses on participants who teach students in grade four and/or five. The policy actually includes all new elementary school teachers and middle school special education teachers, so the results do not include all of the teachers impacted by the policy. Since the policy was enacted recently, most of the effects are felt by participants new to teaching. Inferences about the association between teachers' mathematics content knowledge and their students' performance should therefore be

limited to inexperienced teachers. I also must take into account the internal validity

threats of selection and attrition here as well, since these factors influenced the teacher

participants that were part of the models. The checks on possible attrition effects, for

example, should be factored in to the discussion of the generalizability of results.

The other key external validity threat relevant to this study is *interaction of*

causal relationship with outcomes. As discussed in the construct validity section, the

reliance on MCAS as the sole measure of student outcomes narrows the conclusions I can

draw about student learning. Concerns that MCAS measures student achievement in

ways that may be less rich or holistic than assessments such as portfolios indicate an

external validity threat along the same lines. The use of a student achievement measure

in itself as the only outcome limits what I can say about the impact of the teacher

licensure policy and emergency amendment. There are many other outcomes of the

licensure policy that would be of interest in a full-fledged policy analysis, such as the

impact of the policy on teacher preparation programs or effects on the motivation and

self-confidence of teachers who receive conditional passing scores. Although the focus

of this study on student achievement as measured by MCAS is in line with teacher

effectiveness research, I must be clear that the results do not automatically generalize or

represent a comprehensive policy analysis.

CHAPTER 4

RESULTS

This chapter provides descriptive data about the study sample both for teachers and for students, the results of the main regression analyses, and the results of supplementary analyses, including those related to validity threats.

Descriptive Statistics

Chapter 3 detailed the way in which variables were developed and the state-provided databases were merged. The study sample consisted of 130 teachers and their 2640 students. Table 2 shows the descriptive statistics for teachers in the sample. The sample was comprised of teachers who took the MTEL mathematics subtest prior to May 2010 and were linked in the database with grade 4 and/or grade 5 students who had math SGP scores. Study sample teachers were significantly less experienced with only 33% possessing over two years of experience compared to 89% for the rest of grade 4 and 5 teachers. The study sample group was also about twice as likely to have gained a preliminary license only (21% vs. 10%). The study sample was not significantly different in terms of the percentage of teachers who possessed graduate degrees, mathematics majors, or were classified as special education teachers. Since only study sample teachers were linked to student data, only teacher background data was compared to the full statewide teacher data set.

Table 2

Descriptive Statistics for Teacher Variables, by Inclusion in Study Sample

Variable	Teachers included in full study sample (n = 130)	All other grade 4 and 5 teachers in the state[a] (n = 6,899)
Experience greater than 2 years	.33***	.89
Possesses graduate degree	.58	.66
Holds preliminary license only	.21***	.10
Majored in mathematics	.02	.01
Special education teacher	.13	.18

[a]Teachers included in this comparison had job classification of "teacher" rather than "co-teacher" or "support teacher," corresponding with the selection criteria for the teachers in the study sample.

*p < .05. **p < .01. ***p < .001

All 130 teachers in the study sample were included in the "student-level" regression analyses, but 49 (38%) of these teachers taught fewer than 20 students. Following MADESE guidelines that require at least 20 students in order to attribute median student growth to teacher impact (for the purpose of teacher evaluation), these 49 teachers were removed for the teacher-level analyses that used median student growth as the dependent variable. As shown in table 3, almost all of the teachers removed because they taught fewer than 20 students were special education teachers, so that only 1% of the teachers ultimately included in the teacher-level analyses were special education teachers compared to 33% of the teachers who were not included. The teachers included in the teacher-level analyses were less experienced than those who were excluded and taught significantly lower percentages of special education students, higher percentages of low-

Table 3

Descriptive Statistics for Teacher Variables, by Inclusion in Teacher-Level Regression Analyses

Variable	Teachers included in teacher-level analyses (n = 81)	Teachers not included in teacher level-analyses[a] (n = 49)
Categorical variables		
Experience greater than 2 years	.26*	.45
Possesses graduate degree	.59	.57
Holds preliminary license only	.21	.20
Special education teacher	.01***	.33
Math MTEL pass status	.79	.78
Continuous variables		
Percentage of SPED students	15.34***	49.89
	(14.42)	(43.51)
Percentage of low income students	37.96**	55.92
	(34.50)	(32.04)
Percentage of LEP students	9.13	8.32
	(12.72)	(19.64)
Prior achievement of students[b]	242.54***	232.65
	(10.37)	(11.85)

Note. Standard deviations for continuous variables in parentheses. The p-values were calculated using t-tests for categorical variables and chi-square tests for continuous variables. SPED = special education; LEP = Limited English Proficient.
[a]Teachers not included in teacher-level analyses if they were linked with fewer than 20 students. [b]Average math MCAS score from the prior year for each teacher's students.
*p < .05. **p < .01. ***p < .001

income students. They also taught students who had significantly lower prior achievement. There was no significant difference between included and excluded teachers in terms of pass rates on the MTEL mathematics subtest.

The study sample's 130 teachers represented 1.8% of teachers in the state and their students represented a corresponding 1.8% of the grade 4 and 5 students in the state. As shown in Table 4, the study sample students were significantly more likely to be African-American, Hispanic, and low income than other grade 4 and 5 students. The study sample students had significantly lower prior achievement but there was no significant difference in their growth during the 2009-2010 school year as measured by SGP.

Teachers in the study sample can be split into three major groups, based on the pass status on the MTEL math test: teachers who *fully passed* by meeting the permanent cut score of 240, teachers who *conditionally passed* by meeting the temporary cut score of 227, and teachers who scored below 227 and therefore *failed* to reach even the lower, temporary conditional cut score. Table 5 compares the three groups of teachers and reflects the fact that the three groups did not have significant differences in terms of background variables and the makeup of their classrooms. It is worth noting that for possession of a graduate degree, although it did not emerge in the comparison between these three groups, when the conditional pass and fail groups were combined, their proportion of graduate degree possession of 40% was significantly lower than the 60% of teachers who fully passed, X^2 (1, N = 130) = 5.4, p < .05. Due to the fact that 93% (n = 121) of teachers in the study were white, only 4% (n = 5) were African-American, and none of the teachers were Hispanic, the cell sizes were too small to statistically compare

Table 4

Descriptive Statistics for Student Variables, by Inclusion in Study Sample

Variable	Students included in study sample (n = 2640)	All other grade 4 and 5 students in the state (n = 141,751)
Categorical variables		
African-American	.13***	.08
Hispanic	.20***	.16
Male	.50	.51
Low income	.40**	.37
Limited English Proficient	.07**	.08
Special education	.18	.19
Continuous variables		
Math Student Growth Percentile	49.33	50.10
(SGP)	(29.36)	(28.88)
Prior achievement[a]	241.60***	243.81
	(16.71)	(16.95)

Note. Standard deviations for continuous variables in parentheses. The p-values were calculated using t-tests for categorical variables and chi-square tests for continuous variables.

[a]Student's math MCAS scale score from the prior year.

*p < .05. **p < .01. ***p < .001

teacher race by math MTEL pass status or to include teacher race in the regression models. Other than graduate degree when the fully pass teacher group was compared to all other teachers, then, which mirrors the way the variable was operationalized in the regression analyses, the three teacher groups were effectively equivalent.

Table 5

Descriptive Statistics for Teacher Variables, by Math MTEL Test Pass Status

Variable	Fully passed (n = 102)	Conditionally passed (n = 19)	Failed (n = 9)
Categorical variables			
Experience greater than 2 years	.29	.37	.67
Possesses graduate degree	.64	.47	.22
Holds preliminary license only	.22	.26	.00
Math degree	.02	.00	.00
Special education teacher	.13	.11	.22
Continuous variables			
Percentage of SPED students	29.35	20.34	34.21
	(33.82)	(29.04)	(38.09)
Percentage of low income students	44.55	43.53	49.25
	(35.15)	(31.29)	(37.45)
Percentage of LEP students	9.48	5.08	8.82
	(16.42)	(10.85)	(15.09)
Prior achievement of students[b]	239.09	240.63	232.37
	(22.01)	(8.57)	(15.33)

Note. Standard deviations for continuous variables in parentheses. The p-values were calculated using one-way ANOVA for categorical variables and chi-square tests for continuous variables, with significance tested for conditional pass group compared to the fully pass group and for the failed group compared to the conditional pass group. SPED = special education; LEP = Limited English Proficient.

[a]Average math MCAS score from the prior year for each teacher's students.

*p < .05. **p < .01. ***p < .001

Table 6 displays the descriptive statistics for students by the pass status of their

teachers. The students were quite comparable and showed no overt matching effects in

which teachers who passed the math MTEL were matched up with students who had

background characteristics that were commonly associated with higher test performance.

Students whose teacher conditionally passed were only significantly different than

students whose teacher fully passed in the sense that they were less likely to be African-

Table 6

Descriptive Statistics for Student Variables, by Teacher's Math MTEL Test Pass Status

Variable	Teacher fully passed Math MTEL (n = 2042)	Teacher conditionally passed Math MTEL (n = 395)	Teacher failed Math MTEL (n = 293)
Categorical variables			
White	.56	.59	.71***
African-American	.14	.09**	.11
Hispanic	.20	.23	.06***
Asian	.06	.06	.08
Male	.50	.49	.48
Low income	.40	.43	.25***
Limited English Proficient	.07	.07	.02***
Special education	.19	.12*	.17
Continuous variable			
Prior achievement[a]	241.44	241.12	241.66
	(10.13)	(7.84)	(9.09)

Note. Standard deviations for continuous variable in parentheses. The p-values were calculated using one-way ANOVA for categorical variables and chi-square tests for continuous variables, with significance noted for conditional pass group compared to the fully pass group and for the failed group compared to the conditional pass group.

 [a]Student's math MCAS scale score from the prior year.

*p < .05. **p < .01. ***p < .001

American, with 9% vs. 14% respectively. The rest of the differences between groups

emerged as differences for students whose teacher failed the math MTEL: they were

significantly more likely to be Hispanic, low income, and Limited English Proficient than

students whose teacher conditionally passed. The fact that all of these factors were

potentially favorable for students with teachers who failed and the equivalency of

students' prior achievement across the groups suggests that students were not explicitly

sorted according to teacher pass status.

<p style="text-align:center">Regression Analyses</p>

This section presents the regression models outlined in Chapter 3. The teacher-

level (Model 1) regression analyses are presented first, followed by the student-level

(Model 2) regression analyses.

Teacher-level Models

The teacher-level regression models used the teacher-level dependent variable

of median SGP and included both teacher- and classroom-level independent variables.

Table 7 presents results of model 1a, which included teacher *pass status* on the math

MTEL test as the key independent variable. The model emerged as statistically

significant $F (8, 72) = 3.43$, p < .01, with an R^2 value of .276, accounting for

approximately 28% of the variation in teacher median math SGP. The classroom

characteristics of percentage of students who were low income and LEP were significant,

with negative and positive associations respectively. The strongest predictor of a

teacher's median math SGP was the percentage of his or her students who were low

income, with $B = -.427$ indicating that each additional percent of low income students

predicted almost half a point decrease in the teacher's median math SGP. None of the

teacher characteristics emerged as predictors of median math SGP, with classroom

variables accounting for about 26% of the variation in teacher median math SGP in the

initial block of classroom variables (F (4, 76) = 6.53, p < .001, R^2 = .256) and the teacher

variables only increasing the percentage of variation explained by a statistically

insignificant 2% (ΔR^2 = .02).

Table 8 presents results of model 1b, which included teacher *scores* on the math

Table 7

Model 1a: OLS Regression Analysis of Variables Predicting Teacher Median Math SGP,
Including Teacher Pass Status on Math MTEL Test

Variable	B	SE B	β	95% CI
Classroom variables				
Percentage of SPED students	-.158	.149	-.126	[-.46, .14]
Percentage of low income students	-.427	.100	-.815***	[-.63, -.23]
Percentage of LEP students	.599	.19	.422**	[.22, .98]
Prior achievement[a]	-.554	.319	-.313	[-1.19, .08]
Teacher variables				
Teacher experience greater than 2 yrs	.573	4.360	.014	[-8.12, 9.27]
Teacher possesses graduate degree	2.768	3.871	.076	[-4.95, 10.49]
Teacher holds prelim license only	-3.790	4.629	-.086	[-13.02, 5.44]
Math MTEL pass status	3.648	4.785	.083	[-5.89, 13.19]
R^2		.276		
F		3.428**		

Note. N = 81. CI = confidence interval. SPED = special education. LEP = Limited
English Proficient.
[a]Average math MCAS score from the prior year for each teacher's students.
*p < .05. **p < .01. ***p < .001

MTEL test as the key independent variable. The model emerged as statistically

significant, F (8, 72) = 3.62, p < .01, with an R^2 value of .287, accounting for

approximately 29% of the variation in teacher median math SGP. The significant

variables mirrored the results of model 1a, with percentage of students who are low

Table 8

*Model 1b: OLS Regression Analysis of Variables Predicting Teacher Median Math SGP,
Including Teacher Scale Score on Math MTEL*

Variable	B	SE B	β	95% CI
Classroom variables				
Percentage of SPED students	-.166	.147	-.133	[-.46, .13]
Percentage of low income students	-.432	.099	-.826***	[-.63, -.24]
Percentage of LEP students	.581	.190	.409**	[.20, .96]
Prior achievement	-.591	.315	-.334	[-1.22, .04]
Teacher variables				
Teacher experience greater than 2 years	1.413	4.398	.035	[-7.35, 10.18]
Teacher possesses graduate degree	2.385	3.829	.065	[-5.25, 10.02]
Teacher holds preliminary license only	-2.902	4.617	-.066	[-12.11, 6.30]
Math MTEL scale score	.120	.091	.143	[-.06, .30]
R^2		.287		
F		3.623**		

Note. N = 81. CI = confidence interval. SPED = special education. LEP = Limited English
Proficient.

[a]Prior achievement is defined here as the average math MCAS score from the prior year for each
teacher's students.

*p < .05. **p < .01. ***p < .001

income and percentage of students who are LEP again exhibiting respective positive and

negative associations. Again, no teacher variables were significant predictors of teacher

median math SGP.

Student-level Models

The student-level regression analyses (Model 2) proceeded by adding variables

in four blocks to track how different types of predictors contributed to the full model.

Block one included only student background variables, block two added classroom

variables, block three added teacher characteristics, and block four added teacher results

on the MTEL math subtest. Table 9 shows that student background variables produced a

statistically significant model on their own, $F (6, 2633) = 12.33$, $p < .001$, with an R^2

value of .027. Low income and special education status both were significant predictors,

Table 9

Model 2, Block 1: OLS Regression Analysis of Student Variables Predicting Math SGP

Variable	B	SE B	β	95% CI
African-American	2.020	1.774	.024	[-1.46, 5.49]
Hispanic	1.240	1.672	.017	[-2.04, 4.52]
Male	1.022	1.100	.018	[-1.13, 3.18]
Low income	-8.506	1.344	-.147***	[-11.14, -5.87]
Limited English Proficient	3.707	2.317	.033	[-.84, 8.25]
Special education	-6.753	1.444	-.091***	[-9.59, -3.92]
R^2			.027	
F			12.328***	

Note. N = 2640. CI = confidence interval.

*p < .05. **p < .01. ***p < .001

with Beta values suggesting that being low income was associated with a math SGP about 8.5 points lower than being high income, and a special education student predicted to have a math SGP about 6.8 points lower than a regular education student, when controlling for other student background variables. For comparison purposes, the MADESE suggests that SGPs between 40-60 are "typical growth" (see MADESE, 2011),

Table 10

Model 2, Block 2: OLS Regression Analysis of Student and Classroom Variables Predicting Math SGP

Variable	B	SE B	β	95% CI
Student variables				
African-American	2.669	1.808	.032	[-.88, 6.21]
Hispanic	1.901	1.706	.027	[-1.44, 5.24]
Male	.994	1.100	.018	[-1.16, 3.15]
Low income	-7.881	1.397	-.136***	[-10.62, -5.14]
Limited English Proficient	4.071	2.325	.036	[-.49, 8.63]
Special education	-6.484	1.512	-.087***	[-9.45, -3.52]
Classroom variables				
Class size	-.118	.116	-.020	[-.34, .11]
Peer effect[a]	.124	.068	.042	[-.01, .26]
R^2		.029		
F		9.736***		

Note. $N = 2640$. CI = confidence interval.

[a]Peer effect is defined as the average prior year math MCAS scores of students in the classroom.

*p < .05. **p < .01. ***p < .001

Table 11

Model 2, Block 3: OLS Regression Analysis of Student, Classroom, and Teacher
Variables Predicting Math SGP

Variable	B	SE B	β	95% CI
Student variables				
African-American	2.305	1.807	.028	[-1.24, 5.85]
Hispanic	1.677	1.708	.023	[-1.67, 5.03]
Male	1.005	1.097	.018	[-1.15, 3.16]
Low income	-7.520	1.405	-.130***	[-10.28, -4.76]
Limited English Proficient	4.206	2.320	.037	[-.34, 8.76]
Special education	-6.582	1.514	-.089***	[-9.55, -3.61]
Classroom variables				
Class size	-.076	.120	-.013	[-.31, .16]
Peer effect[a]	.087	.069	.030	[-.05, .22]
Teacher variables				
Experience greater than 2 years	.295	1.223	.005	[-2.1, 2.69]
Possesses graduate degree	4.538	1.123	.079***	[2.34, 6.74]
Holds preliminary license only	-1.270	1.348	-.020	[-3.91, 1.37]
R^2			.035	
F			8.647***	

Note. $N = 2640$. CI = confidence interval.

[a]Peer effect is defined as the average prior year math MCAS scores of students in the classroom.

*p < .05. **p < .01. ***p < .001

so the impact of low income in this case would take a student from an average SGP of 50, almost into the high growth range.

Table 10 reflects the addition of classroom level variables in block two. The model is significant, $F (8, 2631) = 9.74$, p < .001, but the R^2 value of .029 produces an

insignificant ΔR^2 of .002 from block 1. Low income and special education were again negatively associated with student math SGP. Neither of the classroom variables were significant predictors.

Table 11 shows the impact of teacher characteristics in block three. The model is significant, $F(11, 2628) = 8.65$, $p < .001$, and the R^2 value of .035 produced a statistically significant ΔR^2 of .006 from block 2. Among the three teacher variables, only graduate degree was statistically significant ($p < .001$), with a student who was taught by a teacher possessing a graduate degree predicted to have a math SGP about 4.6 points higher than a student taught by a teacher with no graduate degree.

Table 12 shows block four, the full model, in this case adding teacher math MTEL pass status as the key independent variable. The model was significant, $F(12, 2627) = 8.85$, $p < .001$, and the R^2 value of .039 produced a statistically significant ΔR^2 of .004 ($p < .01$) from block 3. The unstandardized Beta estimated that a student had a math SGP about 4.6 points higher if he or she was taught by a teacher who passed the math MTEL test than if he or she was taught by a teacher who did not pass. In terms of standard deviations, based on $SD = 28.351$ for math SGP, the effect size of passing the math MTEL test was .16 standard deviations. The magnitude of association for pass status was larger than the effect of having a teacher who possessed a graduate degree, which was associated with an increase in math SGP of about 3.6 points, and considerably smaller than the estimated impact of being low income or special education, which predicted math SGP scores that were lower by about 7.5 and 6.9 points, respectively. In terms of math SGP points, the benefit of having a teacher who passed the math MTEL made up for about 60% of the negative impact of being low income.

Table 12

Model 2a, Block 4: OLS Regression Analysis of Student, Classroom, and Teacher Variables Predicting Math SGP, Including Teacher Pass Status on Math MTEL Test

Variable	B	SE B	β	95% CI
Student variables				
African-American	1.909	1.808	.023	[-1.64, 5.46]
Hispanic	1.461	1.706	.020	[-1.88, 4.8]
Male	1.013	1.095	.018	[-1.13, 3.16]
Low income	-7.471	1.403	-.129***	[-10.22, -4.72]
Limited English Proficient	3.995	2.317	.035	[-.55, 8.54]
Special education	-6.877	1.514	-.093***	[-9.85, -3.91]
Classroom variables				
Class size	-.033	.121	-.006	[-.27, .2]
Peer effect[a]	.076	.069	.026	[-.06, .21]
Teacher variables				
Experience greater than 2 years	1.140	1.248	.018	[-1.31, 3.59]
Possesses graduate degree	3.594	1.157	.063**	[1.32, 5.86]
Holds preliminary license only	-1.492	1.347	-.023	[-4.13, 1.15]
Math MTEL pass status	4.577	1.395	.068**	[1.84, 7.31]
R^2		.039		
F		8.853***		

Note. $N = 2640$. CI = confidence interval.

[a]Peer effect is defined as the average prior year math MCAS scores of students in the classroom.

*p < .05. **p < .01. ***p < .001

When math MTEL score was used as the key independent variable, the results were quite similar to those for MTEL pass status. As shown in Table 13, the model was significant, $F(12, 2627) = 11.83$, $p < .001$. In this case the R^2 value of .051 reflected a

Table 13

Model 2b, Block 4: OLS Regression Analysis of Student, Classroom, and Teacher Variables Predicting Math SGP, Including Teacher Score on Math MTEL Test

Variable	*B*	SE *B*	β	95% CI
Student variables				
African-American	.667	1.809	.008	[-2.88, 4.21]
Hispanic	.523	1.702	.007	[-2.81, 3.86]
Male	.941	1.088	.017	[-1.19, 3.08]
Low income	-7.433	1.394	-.128***	[-10.17, -4.7]
Limited English Proficient	3.172	2.306	.028	[-1.35, 7.69]
Special education	-7.082	1.503	-.095***	[-10.03, -4.13]
Classroom variables				
Class size	-.066	.119	-.011	[-.30, .17]
Peer effect[a]	.057	.068	.019	[-.08, .19]
Teacher variables				
Experience greater than 2 years	3.281	1.291	.053*	[.75, 5.81]
Possesses graduate degree	2.930	1.139	.051*	[.70, 5.16]
Holds preliminary license only	-.383	1.343	-.006	[-3.02, 2.25
Math MTEL scale score	.180	.027	.143***	[.13, .23]
R^2		.051		
F		11.831***		

Note. $N = 2640$. CI = confidence interval.

[a]Peer effect is defined as the average prior year math MCAS scores of students in the classroom.

*p < .05. **p < .01. ***p < .001

statistically significant ΔR^2 of .006 (p < .01) from block 3. This was the same ΔR^2 as from block 2 to block 3, meaning that math MTEL score provided the same amount of

Table 14

Model 2a for Students Whose Teachers Fully Passed or Conditionally Passed Math MTEL Test

Variable	*B*	SE *B*	β	95% CI
Student variables				
African-American	1.847	1.890	.022	[-1.86, 5.55]
Hispanic	1.191	1.752	.017	[-2.25, 4.63]
Male	.813	1.136	.014***	[-1.42, 3.04]
Low income	-7.768	1.452	-.135***	[-10.62, -4.92]
Limited English Proficient	3.539	2.345	.032	[-1.06, 8.14]
Special education	-7.332	1.571	-.099***	[-10.41, -4.25]
Classroom variables				
Class size	-.008	.128	-.001	[-.26, .24]
Peer effect[a]	.110	.071	.038	[-.03, .25]
Teacher variables				
Experience greater than 2 years	3.194	1.330	.049*	[.59, 5.80]
Possesses graduate degree	3.459	1.185	.060**	[1.14, 5.78]
Holds preliminary license only	-1.983	1.366	-.031	[-4.66, .70]
Math MTEL pass status[b]	1.514	1.562	.020	[-1.55, 4.58]
R^2		.042		
F		8.796***		

Note. N = 2437. Results are for OLS Regression of variables predicting math SGP. CI = confidence interval.

[a]Average prior year math MCAS scores of students in the classroom. [b]In this table "pass status" contrasts teachers who fully passed versus those who conditionally passed.

*p < .05. **p < .01. ***p < .001

predictive power as the other three teacher variables combined. Interpreting the

unstandardized Beta yielded an estimate that a one point increase in math MTEL score

was associated with a .18 point increase in math SGP. In this model, teacher experience was a significant predictor along with graduate degree, both at the p < .05 level.

In order to look more closely at the cut points for the math MTEL test, it is instructive to examine the full model as applied to the three categories of teacher pass status. Instead of lumping together students whose teachers either conditionally passed or failed the math MTEL, Table 14 and 15 present regression analyses for subgroups of the study sample as defined by teacher pass status. Table 14 shows that when students whose teachers failed the test were removed from the analyses, the status of teachers as either fully or conditionally passing was not significantly associated with student growth. In contrast, Table 15 reflects a significant difference in growth between students whose teacher conditionally passed the math MTEL compared to students whose teacher failed the math MTEL. The model was significant, $F (12, 585) = 6.799$, p < .001, with an R^2 value of .122. This means that the amount of variation explained doubled from almost 6% to more than 12% by introducing the pass status variable (from block 3 to block 4, $\Delta R^2 = .064$, p > .001). The standardized Beta showed pass status to be the strongest predictor in the entire model. The unstandardized Beta estimated that having a teacher who conditionally passed the math MTEL test was associated with a 23 point higher math SGP than having a teacher who failed the math MTEL, which is about 5 times more than the corresponding estimate from the full study sample. Based on a math MSGP standard deviation of 29.086, this means that the effect size of conditionally passing versus failing the math MTEL test was .79 standard deviations. In addition to and while controlling for pass status, students who were Limited English Proficient and whose teachers were more experienced were predicted to have higher growth. Students who were low income,

Table 15

Model 2a for Students Whose Teachers Conditionally Passed or Failed Math MTEL Test

Variable	*B*	SE *B*	β	95% CI
Student variables				
African-American	-.622	4.082	-.006	[-8.64, 7.40]
Hispanic	-6.471	3.592	-.085	[-13.53, .58]
Male	.517	2.288	.009	[-3.98, 5.01]
Low income	-10.603	2.907	-.176***	[-16.31, -4.89]
Limited English Proficient	11.021	5.393	.084*	[.43, 21.61]
Special education	-.783	3.443	-.009	[-7.55, 5.98]
Classroom variables				
Class size	.018	.290	.003	[-.55, .59]
Peer effect[a]	-.796	.190	-.227***	[-1.17, -.42]
Teacher variables				
Experience greater than 2 years	14.500	3.369	.249***	[7.88, 21.12]
Possesses graduate degree	.677	2.818	.011	[-4.86, 6.21]
Holds preliminary license only	-13.122	3.824	-.183**	[-20.63, -5.61]
Math MTEL pass status[b]	23.060	3.525	.376***	[16.14, 29.98]
R^2		.122		
F		6.799***		

Note. $N = 598$. Results are for OLS Regression of variables predicting math SGP. CI = confidence interval.

[a]Average prior year math MCAS scores of students in the classroom. [b]In this table "pass status" contrasts teachers who conditionally passed versus those who failed.

*p < .05. **p < .01. ***p < .001

whose peers' prior achievement was lower, and whose teachers possessed preliminary

licensure only were predicted to have lower growth. While more factors were significant

than prior models, the key finding here was the statistically significant, relatively large, and practically meaningful predictive power of knowing whether a student's teacher scored above or below the conditional cut score.

Supplementary Analyses

This section presents further analyses that are relevant to validity threats, first looking at regression models that use English Language Arts (ELA) SGP as the dependent variable in order to allow consideration of the extent to which the results are content-specific, and then exploring other models that shed light on the nature of the main dependent variable used throughout the study, math SGP.

Content-specificity of results

In order to look at whether the association between teacher performance on the math MTEL test and student performance was content-specific, regression analyses were conducted that used student growth in ELA as the outcome measure. Table 16 shows that when using math MTEL test pass status as the key independent variable, although the model was significant, $F(12, 2657) = 9.79$, p < .001, math MTEL pass status was not a predictor of ELA SGP. Table 17 shows that math MTEL test score produced an association with ELA SGP when used as the key independent variable, though it was barely significant with the 95% confidence interval starting as close to 0 as possible, at .01. The unstandardized Beta estimated an impact of .05 SGP points per additional scale score point on the math MTEL, compared to .18 for this model when math SGP was used. The math MTEL test was a substantially weaker predictor of ELA growth than it was of math growth, suggesting that there was a content-specific aspect to the association between teacher test score and student performance.

118

Table 16

OLS Regression Analysis of Variables Predicting ELA SGP, Including Teacher Pass Status on Math MTEL Test

Variable	*B*	SE *B*	β	95% CI
Student variables				
African-American	3.596	1.850	.042	[-.03, 7.22]
Hispanic	-.120	1.734	-.002	[-3.52, 3.28]
Male	-6.882	1.116	-.118***	[-9.07, -4.69]
Low income	-4.567	1.436	-.077**	[-7.38, -1.75]
Limited English Proficient	.199	2.289	.002	[-4.29, 4.69]
Special Education	-5.851	1.552	-.076***	[-8.89, -2.81]
Classroom variables				
Class size	-.123	.124	-.021	[-.37, .12]
Peer effect[a]	.175	.069	.060*	[.04, .31]
Teacher variables				
Experience greater than 2 years	1.641	1.273	.026	[-.86, 4.14]
Possesses graduate degree	2.696	1.182	.046*	[.38, 5.01]
Holds preliminary license only	-.606	1.365	-.009	[-3.28, 2.07]
Math MTEL pass status	2.297	1.416	.033	[-.48, 5.07]
R^2		.042		
F		9.788***		

Note. N = 2670. ELA = English Language Arts. CI = confidence interval.

[a]Average prior year math MCAS scores of students in the classroom

*p < .05. **p < .01. ***p < .001

Math SGP as the outcome measure

As the main outcome measure of this study, it is important to examine math SGP in terms of its robustness as a growth measure by conducting additional regression analyses and correlations. One of the main advantages that was assumed based on the

Table 17

OLS Regression Analysis of Variables Predicting ELA SGP, Including Teacher Score on Math MTEL Test

Variable	*B*	SE *B*	β	95% CI
Student variables				
African-American	3.309	1.862	.038	[-.34, 6.96]
Hispanic	-.353	1.741	-.005	[-3.77, 3.06]
Male	-6.904	1.116	-.119***	[-9.09, -4.72]
Low income	-4.571	1.436	-.077**	[-7.39, -1.76]
Limited English Proficient	-.004	2.292	.000	[-4.50, 4.49]
Special education	-5.866	1.551	-.076***	[-8.91, -2.83]
Classroom variables				
Class size	-.141	.124	-.024	[-.38, .10]
Peer effect[a]	.171	.069	.059*	[.04, .31]
Teacher variables				
Experience greater than 2 years	2.111	1.324	.033	[-.48, 4.71]
Possesses graduate degree	2.667	1.172	.045*	[.37, 4.97]
Holds preliminary license only	-.232	1.367	-.003	[-2.91, 2.45]
Math MTEL scale score	.054	.027	.042*	[.01, .11]
R^2		.043		
F		9.905***		

Note. N = 2670. ELA = English Language Arts; CI = confidence interval.

[a]Average prior year math MCAS scores of students in the classroom

*p < .05. **p < .01. ***p < .001

literature is that math SGP was advantageous over an achievement measure such as scaled score because it compared students to their academic peers, thus controlling more robustly for unmeasured characteristics and minimizing its association with student background. Table 18 displays results of a regression analysis that used math MCAS

Table 18

OLS Regression Analysis of Variables Predicting Math MCAS Score

Variable	B	SE B	β	95% CI
Student variables				
African-American	-1.287	.716	-.025	[-2.69, .12]
Hispanic	-1.124	.670	-.025	[-2.44, .19]
Male	.865	.430	.025*	[.02, 1.71]
Low income	-4.614	.553	-.128***	[-5.70, -3.53]
Limited English Proficient	-.909	.907	-.013	[-2.69, .87]
Special education	-5.275	.625	-.115***	[-6.50, -4.05]
Prior achievement[a]	.705	.017	.672***	[.67, .74]
Classroom variables				
Class size	-.004	.047	-.001	[-.09, .09]
Peer effect[b]	-.035	.030	-.019	[-.09, .02]
Teacher variables				
Experience greater than 2 years	1.136	.509	.029*	[.14, 2.14]
Possesses graduate degree	1.639	.450	.046***	[.76, 2.52]
Holds preliminary license only	2.720	.530	.067***	[1.68, 3.76]
Math MTEL scale score	.064	.011	.082***	[.04, .09]
R^2		.61		
F		323.92***		

Note. $N = 2666$. CI = confidence interval.

[a]Prior year math MCAS score. [b]Average prior year math MCAS scores of students in the classroom.

*p < .05. **p < .01. ***p < .001

scaled score as the dependent variable and included students' scores from the prior year

as a representation of a more typical value-added methodology. The model was

significant, $F(12, 2809) = 323.92$, p < .001, and produced seven significant factors (not

counting the prior achievement variable) compared to five significant factors in the

analysis with math SGP (see Table 13 for comparable main analysis). The full model

explained about 61% of the variability in student math MCAS scores, which was more

than 12 times the 5% of the variability in student math SGP explained by the

corresponding main analysis. Yet 60% of the variability was explained by student

characteristics alone. When the math MTEL score was added as a predictor, the ΔR^2 of

.005 (p < .001) from .608 to .614 represented an increase of less than 1% of explained

variability. In contrast, the addition of math MTEL score to the original model produced

a ΔR^2 of .016 (p < .001), from .035 to .051, increasing the explained variability by 45.7%.

The original model also estimated a relatively larger impact, with math MTEL score

associated with a 1.4 standard deviation increase in math SGP versus a .08 standard

deviation increase in math MCAS score in the comparison model. It appears that the use

of SGP as the outcome measure provided models that were less explanatory of variation

in student growth overall but more sensitive to the key independent variable.

Part of the reason why SGP was more sensitive to the impact of the key

independent variable may be because it was less strongly related to student background

and other control variables. Table 19 shows that the correlations of math SGP to student

background were significant yet relatively moderate, $r(2638) = -.133$, p < .001, for low

income and $r(2638) = -.098$, p < .001, for special education status. The correlation

coefficients for math MCAS score were many times larger, $r(2638) = -.382$, p < .001, for

low income and and $r(2638) = -.373$, p < .001, for special education status. Another

important factor is that math SGP was not significantly or strongly related to the prior

year math SGP, $r(1384) = .009$, p = n.s, which suggests that there are not "high growth" students who will post high SGP scores regardless of teacher impact.

Table 19

Correlations Related to Student Outcome Measure

	Math SGP	Math SGP prior year	Math MCAS score	Low income	Special education
Math SGP	____				
Math SGP prior year	.009 (n = 1386)	____			
Math MCAS score	.568*** (n = 2640)	.292*** (n = 1386)	____		
Low income	-.133*** (n = 2640)	-.093** (n = 1386)	-.382*** (n = 2640)	____	
Special education	-.098*** (n = 2640)	-.123*** (n = 1386)	-.373*** (n = 2640)	.086*** (n = 2640)	____

* p < .05. ** p < .01. *** p < .001

CHAPTER 5

DISCUSSION

The results reject the null hypothesis and provide evidence that there is an association between teacher performance on the MTEL math test and student growth measured by SGP. This chapter summarizes the study findings, discusses implications, and provides suggestions for further research.

Summary of the Findings

Research question and key independent variable

In response to the overarching research question, this study provides evidence that the math MTEL test is a valid predictor of teacher effectiveness. The primary subquestion is whether teacher outcomes on the math MTEL test are associated with student growth on MCAS, as measured by math SGP. This question effectively defines predictive validity from the standpoint of standards-based education reform. Based on these two standardized and state-provided instruments, in this study a student's predicted growth in mathematics rose in accordance with their teacher's score on the licensure test. Correspondingly, if a student had a teacher who had fully passed the licensure test, their growth score was predicted to be higher than a student who had a teacher who had conditionally passed or failed the licensure test.

Interestingly, when the effects of pass status were parsed between the three groups, the impact of pass status seems to have come from the difference between

124

teachers who failed the math MTEL test and teachers who conditionally passed. Although no difference in predicted student growth was found when looking at students whose teachers fully passed the math MTEL test versus students whose teachers conditionally passed, there were strong results when analyses were limited to the failed and conditionally passed teacher groups. Although the sample size was less than a quarter of the size of the main regression analyses, the results from the parsed analyses suggest that the predictive validity of the math MTEL test is solid at the cut score of 227 set by the emergency amendment and that evidence is considerably weaker for predictive validity based on determinations made using the permanent cut score of 240.

The strength of the relationship found in this study is considerably higher than what is found in the literature. Goldhaber and Hansen (2009), the only study to measure effect sizes in terms of teacher score on the licensure test, estimated that a one standard deviation increase in teacher test score was associated with an increase in grade 4-6 student math growth of .015 standard deviations, which is nearly 10 times smaller than the .143 effect size found here. In terms of pass status, Goldhaber and Hansen (2009) estimated an effect size of .05, along the lines of what Clotfelter et al. (2010) found for high school algebra but much larger than the .015 estimated by Clotfelter et al. (2007) for grade 3-5 students. The main regression analyses in this study estimated the effect size of pass status as .16 standard deviations, triple the high end from the literature.

When looking at the difference between students whose teachers conditionally passed and students whose teachers failed, the effect size of pass status jumps to a relatively huge .79 standard deviation units. Considering that the MADESE recommends that math SGP scores between 40 and 60 points be interpreted as "typical growth," the

associated difference of 23 math SGP points suggests that a student being placed with a teacher who conditionally passed versus failed the math MTEL test could make the difference between low growth and high growth in a year of math learning. The results of this study are therefore large relative to the literature and are practically significant as well.

Research subquestions and other independent variables

The second research subquestion looks at the strength of the association between teacher licensure test outcomes and student growth in comparison to other teacher characteristics. In the main regression analyses, graduate degree showed significant impacts along with teacher pass status, while experience and preliminary license were unassociated with student growth. As dummy variables, graduate degree and teacher pass status are directly comparable and showed similar levels of impact. In the main model that used teacher test score as the key independent variable, both graduate degree and experience emerged as significant predictors. In that case the addition of teacher test score increased the amount of variability explained by the model as much as the addition of the rest of the teacher characteristics.

The finding that the predictive power of teachers' performance on the licensure test was on par with or exceeded other teacher characteristics in this study is a divergence from the literature, in which the effects of teacher licensure test performance has been a very small portion of the package of teacher credentials (Boyd, et al., 2007; Clotfelter, et al., 2007; Clotfelter, et al., 2010; Goldhaber & Hansen, 2009; Wayne & Youngs, 2003) or nonexistent (Buddin & Zamarro, 2009). Clotfelter et al. (2007), for example, showed the effects of teacher experience to be nearly seven times larger than teacher licensure test

126

pass status (Clotfelter, et al., 2007). The results of this study suggest that policymakers get substantial information from teacher outcomes on the math MTEL test that they would not get from looking at other known teacher characteristics alone.

The final research subquestion involves student characteristics as the comparative benchmark for the impact of teacher licensure outcomes. Here, too, I found that the strength of the association of teacher licensure outcomes stand up fairly well. In the main regression analyses the student characteristics of low-income status and special education status exhibited significant negative associations with student growth in mathematics. For teacher pass status a student's math SGP was predicted to be 4.58 points higher when their teacher had passed the math MTEL test, which is about 60% of the predicted decrease of 7.47 points from being low income and two thirds of the predicted decrease of 6.88 points from special education status. These results are not far off from the relative effect sizes for pedagogical content knowledge in mathematics that were hailed as hugely significant when on par with SES (Hill, Rowan, & Ball, 2005).

In the regression analyses that looked at teachers in the conditional pass versus failed groups, the relative effect of math MTEL pass status was much more impressive, with the gain from having a teacher who conditionally passed estimated at over twice the size of the detriment of low income status. The results for student characteristics for that model were inconsistent with the main regressions, however, with special education status showing no association and Limited English Proficient status emerging as a positive predictor. Keeping in mind the fact that the sample size was a quarter of the size as the main regression analyses, I should interpret the specific variable-by-variable results

cautiously. On the whole, though, this study provided evidence that the impact of the math MTEL test on student growth would not be washed away by student background.

Without over-interpreting results for the other independent variables, since they were primarily included as covariates in the overall model in order to support inferences about the association between teacher test score and student growth, I can say that the results generally align with the literature. Teacher experience was found to be a significant predictor, which is strongly supported in the literature within the first few years of teaching (Clotfelter, et al., 2007; Harris & Rutledge, 2010; Kukla-Acavedo, 2009; Rice, 2003). Although teacher experience enjoys perhaps the most consistent overall support in the literature as a teacher characteristic, the association with elementary students' math performance here is notable, since some studies have suggested that experience matters less at the elementary level than the secondary level (Buddin & Zamarro, 2009; Jepsen, 2005) and less in math than reading (Croninger, et al., 2007; Rockoff, 2004).

Graduate degree was a significant predictor of student outcomes, which has been somewhat inconsistent in the literature according to reviewers (Goe, 2007; Kennedy, Ahn, & Choi, 2008; Wayne & Youngs, 2003), but has shown up as a factor more strongly in mathematics (Cochran-Smith & Zeichner, 2005; Goldhaber & Brewer, 1997, Goldhaber & Brewer, 2000; Rice, 2003) and at the elementary level (Harris & Sass, 2007).

Teacher licensure did not emerge in the main regression analyses as a predictor, though it did show up in the smaller sample size used for the conditional pass versus fail regression analyses, and this matches the famous debate on licensure in which

researchers preferring large scale teacher effectiveness studies argued that licensure does not matter (Goldhaber & Brewer, 2000), while those who give more credence to smaller scale and qualitative studies continued to assert that it does (Darling-Hammond, Berry, & Thoreson, 2001). The results here match the conclusions of previous reviewers who found the evidence for licensure to be mixed (Cochran-Smith & Zeichner, 2005; Goe, 2007; Rice, 2003; Wayne & Youngs, 2003).

Although the literature has shown mixed results for the effects of teacher race, teacher gender, math-specific degree, and teacher cognitive ability, none of these variables were included due to lack of variation in the data. Output from regression analyses that include math degree and teacher score on the MTEL Communications and Literacy Test (as a proxy for teacher cognitive ability) are included in Appendix A; the results mirrored the main regression analyses.

Overall the models in this study were less explanatory that those in the literature. The main regression analyses in this study reflected 5-12% of explained variation compared to 60-70% in studies that used similar sets of variables (Clotfelter et al., 2007; Goldhaber & Hansen, 2009). This discrepancy is likely due to the fact that math SGP is less strongly associated with the control variables than other growth measures. When the models were applied using a more traditional growth measure, with MCAS score as the dependent variable along with a control variable for prior achievement, the percent of variability explained shot up to 61%. Table 18 shows that when MCAS was used as the dependent variable, more covariates emerged as a significant predictor and the estimated impact of math MTEL score were lower, even though the additional proportion of variability explained by adding the key independent variable were much higher. Since

this study was focused on one variable, rather than emphasizing comparisons between variables or explaining as much variation as possible, the apparent trade-off for increased sensitivity to the variable of interest can be considered a good one.

Validity

I will consider each of the four categories of validity derived from Shadish, Cook, and Campbell (2002) and discussed in Chapter 2.

Statistical conclusion validity. It was established that the sample size required to detect a small effect size was 869 and the sample size required for a medium effect was 29. The student-level models were evidently sufficiently powered to detect associations. The teacher models in this study were based on 81 teachers. The effect sizes found for the student-level results suggest that the effects were at least in the medium category, but this may not translate directly to effect sizes at the teacher-level. Although it is therefore possible that the teacher-level models were underpowered, those models serve the purpose of informing implications in terms of the dilemma of attributing student results to individual teachers in contexts such as educator evaluation.

One potential statistical issue for this study is that math SGP is uniformly distributed, rather than normally distributed which is an assumption of OLS regression techniques. In order to test for this, math SGP scores were converted to Normal Curve Equivalents (NCEs) and then the regression analyses were conducted using NCEs as the outcome variable. As shown in Table Appendix B, the results were nearly equivalent, suggesting that the lack of normality in math SGP was not a substantial problem. The estimated impact of the math MTEL test was very slightly lower in the NCE model,

suggesting the possibility of a small upward bias in the main regression analyses. Math SGP was used in the main regression analyses for ease of interpretation.

To examine the validity threat of unmeasured variables, I examined the study sample and the dependent variable for evidence that students and teachers were matched in a way that varied systematically alongside teacher outcomes on the math MTEL test. When the three different groups of teachers (based on pass status) were compared, no differences emerged in terms of background, percentage of students of each category (low income, SPED, LEP) that they taught, or the prior achievement of their students (see Table 5). At the student level, the differences were the opposite of what would be expected if teachers with better outcomes on the math MTEL test were being matched with more favorable students: teachers who failed taught students who were proportionally more White and less likely to be Hispanic, low income, and LEP (see Table 6). Based on the background variables of teachers, classrooms, and students, there is no evidence for nonrandom sorting that would bias results.

The other aspect of this study that mitigates against unmeasured variables is the nature of the dependent variable. One of the main unmeasured variables that could confound results is the internal motivation of students. If students who are highly motivated are systematically sorted into classrooms taught by teachers who performed well on the math MTEL test, then I may falsely conclude that teacher test performance is a predictor of student growth when in fact the students' motivation is the key factor. Students with high internal motivation would likely be "high growth" students as well, which would show up as a correlation in SGP scores from one year to the next. In the sample I could look at this possibility for grade 5 students (since students first receive an

SGP score at grade 4), and Table 19 shows a small and insignificant correlation between students' 2011 math SGP and 2010 math SGP. This would suggest that math SGP reflects the growth that a student makes within a given year, and that the cumulative effects of many unmeasured variables such as internal motivation and home life are captured due to the fact that SGP is calculated as a comparison with academic peers. As a measure of student growth, in this study math SGP may have helped guard against nonrandom sorting.

Internal validity. As discussed in Chapter 3, the nature of this study as a natural experiment guards against most of the selection issues that plague other quasi-experiments. Participants in this study were placed into groups (conditions) exogenously based on their teacher test results and did not have the opportunity to self-select their study participation or their condition. When the population of interest is defined as teachers who took the math MTEL test and were now teaching, this study included almost all of those teachers in the student-level models. In the teacher-level models, the 81 teachers who had sufficient numbers of students to be included were less experienced and were general education teachers rather than special education teachers (see Table 3), so there were some possible selection issues, but it is unclear how it might have biased results since there were no differences in student growth between these two groups.

When all teachers who took the math MTEL test are considered, attrition becomes an issue. Publicly available data from MADESE suggests that over 3000 people took the math MTEL test before the start of the 2010-11 school year. Among this larger pool are prospective teachers who are still completing preparation programs and others who decided not to teach or could not secure a teaching position. The data set did allow

us to examine teachers who did not show up in the teaching force whatsoever, so I had no way to explore whether they varied systematically based on their results on the math MTEL test. This issue mirrors the crucial problem that has historically plagued teacher licensure test research, in which teachers who fail the test are not given licenses to teach and therefore do not show up in classrooms. This study suffers from this challenge less than other studies due to the policy that provided temporary licenses for teachers who passed a lower conditional cut score, but there was still a possible attrition effect based on personal decision making by test takers. The directionality of bias from attrition is unclear.

Construct validity. The extent to which the math MTEL test is a valid representation of teacher knowledge in mathematics is unclear. Including a covariate for cognitive ability would have helped the claim that the math MTEL test is specific to mathematical knowledge, but the variable of teacher score on the MTEL Communications and Literacy Test was discarded due to the fact that 95% of teachers had passed the test (which disqualified it as a useful discriminator). Alternatively, the variable of mathematics degree could have been compared to math MTEL results to try to triangulate the construct of mathematics knowledge, but it was left out for similar reasons.

In order to check the content-specificity of results, supplementary regression analyses were performed using ELA SGP rather than math SGP as the dependent variable. The fact that the math MTEL test was a substantially weaker predictor of student growth in ELA than student growth (see Tables 16 and 17) lends some support to the notion that the association between teacher licensure test outcomes and student

growth are due to teacher mathematics knowledge. Yet the literature generally shows teacher credentials to make a larger difference in student mathematics performance than for other subjects. Therefore, the results of this study do not shed light on whether the math MTEL test represents teacher mathematical knowledge or something else such as cognitive ability.

Even with a conservative assumption about what the math MTEL test represents, I must be careful to consider the level at which the construct operated in this study. Recall that the regression analyses comparing teachers who failed to teachers who conditionally passed exhibited much stronger results for math MTEL outcomes than the regression analyses that compared teachers who fully passed to those who conditionally passed. Whatever math MTEL test performance represents, whether it is mathematical knowledge, cognitive ability, or something else, the results of this study apply most clearly to teachers who scored in the lower range.

External validity. The generalizability of results for this study are limited by the fact that the study sample was comprised of teachers and their grade 4 and/or 5 students. These results would not necessarily extend to different grade levels. Further, unsurprisingly since all participants had recently taken a licensure test, teachers in the study sample were significantly less experienced and more likely to have entry-level licenses than other teachers in grade 4 and 5 (see Table 2). As the literature would predict for newer and less credentialed teachers (for example, see Kennedy, Ahn, & Choi, 2008), in comparison to other grade 4 and 5 the student sample was proportionally more African-American, Hispanic, and low income (see Table 4), suggesting that that the teachers in the study sample were more likely to be working in urban schools as well.

Interestingly, although the study sample students had lower prior achievement, their growth was equivalent to students in the rest of the state. Study results generalize most directly, therefore, to teachers entering the profession and teachers serving diverse populations; caution should be employed when interpreting implications for all teachers.

The other major factor to consider is the limitations on inferences that can be made due to the nature of the measures. Specifically, the key independent variable of teacher math MTEL test performance and the main dependent variable of student math SGP are both based on standardized tests. The results may not generalize to relationships where teacher knowledge or student growth are measured in different ways, for example by portfolio assessments or incorporating nonacademic factors such as socio-emotional skills. Strictly speaking, the results can be taken at face value to show that teachers who do well on this particular licensure test are likely to be matched with students who exhibit growth on MCAS as measured specifically by math SGP. This is enough to provide empirical evidence of predictive validity for the test, which is important from a policy perspective. Further policy implications are discussed in the next section.

Policy Implications

Even with possible upward biases taken into account, the strength of the results in this study relative to the literature provides provides firm evidence that the math MTEL test has predictive validity. This is crucial information that augments the content validity that appears to exist due to the robust test development process followed by MADESE. Without predictive validity, there are no assurances that a licensure test is acting as a valid gate-keeper into the teaching profession. The results of this study suggest that the predictive validity of the test is strongest at the lower end of the score

spectrum. An argument could be made that the cut score set by the emergency amendment was the right one, with major differences in predicted student performance based on whether teachers fell on one side (conditional pass) or the other (fail) of that score. The fact that the passing line has automatically, and perhaps arbitrarily, reverted to the original cut score greatly increases the likelihood that licenses are being withheld from teachers unfairly. This study has allowed us to look at those teachers who landed in the conditional pass category and were allowed to teach, ultimately finding little evidence that their students grew less than students of teachers who fully passed. If these teachers are denied licenses and leave the teaching force, the overall quality of teaching could be weakened and student achievement decreased. Based on the results of this study, policymakers should consider returning the cut score to the lower level of 227.

This study provides an important data point to inform the ongoing debate between professionalizers and deregulators. If the results had supported the null hypothesis, showing no predictive validity for the math MTEL test, it would have been a clear win for professionalization advocates. If the teacher test worked against the goal of increasing student achievement, its value would have been undermined based on the standards-based reform logic relied upon by deregulators.

The debate will certainly rage on, however. Predictive validity is a necessary but far from sufficient condition for showing that teacher licensure testing is a component of sound education policy. Professionalization advocates could agree that teacher knowledge is a key aspect of teacher effectiveness but not cede the point that the math MTEL test measures teacher knowledge. A cynical interpretation of these results would be that good test takers are better at teaching students how to take tests. Many of the so-

called unintended consequences of high-stakes student testing would still apply, such as narrowing the curriculum, de-emphasizing critical thinking skills, and dehumanizing the teaching and learning process. To the extent that the math MTEL test is a limited construct of teacher knowledge, and because of the standardized nature of the testing instruments, this study leaves many issues in this debate untouched. I would recommend that both sides of the debate acknowledge that the predictive validity of testing instruments is an area that deserves a lot more attention and that in this case the state should be lauded for trying to incorporate the full research base about teacher knowledge into the test development process.

The results of this study should inform ongoing discussions about another aspect of teacher quality policy, educator evaluation. The teacher-level regression analyses used the same measure of average student growth, median math SGP, which will be used for individual teachers as part of their formal evaluations starting in school year 2014-15, yet these models were not sensitive enough to detect any impact of teacher credentials. It is possible that with a larger sample of teachers, the power of the model would have been sufficient to explain variation between teachers. Yet if these models that include controls did not find associations in a sample size of 81 teachers, it calls into question the validity of equating an individual teacher's median SGP with their effectiveness.

The large number of teachers initially disqualified from this study because they were classified as co-teachers or support teachers, or the 38% of the full study sample ruled out of the teacher-level regressions because they had too few students, underscores other limitations of measuring teacher impact on student growth. Besides the more than 90% of the teachers in the state who work primarily in untested subjects, this study shows

that attributions of teacher impact are difficult to establish at the teacher level even when

robust standardized measures are used. It is unclear how the state will deal with

attributing impact to teachers who share students, or how local schools and districts can

be expected to create measurement instruments reliable enough for making attributions.

This study raises a lot of questions about the pure logistics of including student growth as

a mandatory part of the educator evaluation system.

There are major equity concerns raised by this study as well. The teacher-level

models suggest that the average growth of a teacher's students was negatively associated

with the percentage of those students who are low income (see Table 17 and Table 18).

The student-level models showed a negative association between student growth and

their status as low income or disabled (see Table 12 and Table 13). Although math SGP

was less strongly correlated with these student background variables than math MCAS

scores (see Table 19), the correlations were statistically significant. When SGP was

introduced, MADESE documentation noted "it is not so clear that low income students

should *grow* slower once you've taken performance level into account, given the way we

calculate growth" (MADESE, 2011, p. 19). There was a clear association in this study

and assumedly in MADESE's own analyses as well. Policymakers should initiate a

transparent and genuine conversation about the equity implications of relying on

measures such as SGP for educator evaluation and accountability measures. Otherwise

there are clear incentives for stakeholders, including districts and school leaders as well

as individual teachers, to avoid working with the students from vulnerable populations.

Recommendations for future research

The Massachusetts data set appears to provide a valuable addition to the few data sets presently available to teacher effectiveness researchers. More studies should be conducted using this data set, similar to the manner in which multiple studies by various researchers have employed the North Carolina statewide database. Students and teachers have now been linked for three years and it is likely that the data rules are solidifying and the collection processes improving so that overall it is a cleaner set of data. The statewide data set should be used extensively to explore various questions beyond teacher effectiveness for which standardized teacher-student data are useful.

For future studies looking at predictive validity of the new MTEL math test, now that the data set is longitudinal, research should be conducted that takes advantage of multiple years of data by employing fixed effects to further guard against nonrandom sorting of students and teachers. As the data set grows, larger samples will allow further exploration of covariates as well as examination of whether the predictive power of the math MTEL test vary by teacher race or gender (see Goldhaber & Hansen, 2009). Future studies could take advantage of the changing policy environment by, for example, looking at the effectiveness of teachers who received temporary licenses by conditionally passing but who have yet to reach the full pass score. Finally, in order to answer the question of what exactly the math MTEL test measures, comparative validity studies should be conducted by comparing results on the math MTEL test with results from a validated instrument measuring an established construct of interest, such as the Mathematical Knowledge for Teaching assessments of pedagogical content knowledge (see Hill, Schilling, & Ball, 2004).

Future researchers should conduct thorough policy analyses that examine the impact of the new elementary teacher math subtest and consider various impacts of "raising the bar" of entry in mathematics for elementary and special education teachers. What are the equity impacts of this policy and other teacher licensure requirements? How does the new policy influence schools of teacher education, in terms of mathematical and other coursework, expertise of faculty, standards for admitting students, and graduation requirements? How are teacher candidates affected, practically through their efforts to meet licensure requirements and psychologically through their perceptions of how their competence is measured? How does the implementation of the policy promote or inhibit school improvement efforts by various stakeholders? These analyses should incorporate the results of this study and go beyond purely quantitative methods in order to inform the broader debate about teacher effectiveness and education reform.

APPENDIX A
OLS REGRESSION ANALYSIS OF ALL VARIABLES PREDICTING MATH SGP

Variable	B	SE B	β	95% CI
Student variables				
African-American	.937	1.822	.011	[-2.64, 4.51]
Hispanic	.772	1.723	.011	[-2.61, 4.15]
Asian	10.856	2.254	.094***	[6.44, 15.28]
Male	-7.025	1.394	-.121	[-9.76, -4.29]
Low income	.753	1.083	.013***	[-1.37, 2.88]
Limited English Proficient	1.691	2.304	.015	[-2.83, 6.21]
Special education	-7.478	1.503	-.101***	[-10.43, -4.53]
Classroom variables				
Class size	-.127	.120	-.022	[-.36, .11]
Peer effect[a]	.067	.068	.023	[-.07, .20]
Teacher variables				
Experience greater than 2 years	2.904	1.298	.046*	[.40, 5.45]
Possesses graduate degree	4.733	1.190	.083***	[2.40, 7.07]
Holds preliminary license only	1.214	1.379	.019	[-1.49, 3.92]
Possesses math-specific degree	1.908	4.477	.009	[-6.88, 10.69]
Math MTEL scale score	.229	.033	.176***	[.16, .29]
MTEL Communications and Literacy score	-.208	.045	-.116***	[-.30, -.12]
R^2		.067		
F		12.421***		

Note. $N = 2614$. SGP = Student Growth Percentile. CI = confidence interval. MTEL = Massachusetts Test for Educator Licensure.

[a]Peer effect is defined as the average prior year math MCAS scores of students in the classroom.

*p < .05. **p < .01. ***p < .001

APPENDIX B
OLS REGRESSION ANALYSIS OF VARIABLES PREDICTING
NORMAL CURVE EQUIVALENT OF MATH SGP

Variable	B	SE B	β	95% CI
Student				
African-American	.603	1.280	.010	[-1.91, 3.11]
Hispanic	.473	1.205	.009	[-1.89, 2.84]
Male	.870	.770	.022	[-.64, 1.57]
Free and reduced lunch	-5.186	.986	-.127***	[-7.12, -3.25]
Limited English Proficient	2.251	1.632	.028	[-.95, 5.45]
Disability	-5.005	1.064	-.095***	[-7.09, -2.92]
Classroom Variables				
Class size	-.018	.084	-.005	[-.18, .15]
Prior achievement	.049	.048	.024	[-.05, .14]
Teacher Variables				
Experience greater than 2 years	2.165	.914	.049*	[.37, 3.96]
Possesses graduate degree	1.994	.806	.049*	[.41, 3.58]
Holds preliminary license only	-.234	.950	-.005	[-2.10, 1.63]
Math MTEL scale score	.121	.019	.135***	[.08, .16]
R^2		.045		
F		11.414***		

Note. $N = 2670$. CI = confidence interval.

*p < .05. **p < .01. ***p < .001

REFERENCE LIST

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

Allison, P. D. (2009). *Fixed Effects Regression Models*. Thousand Oaks: Sage Publications, Inc.

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

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

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Ball, D. L., & Forzani, F. M. (2009). The work of teaching and the challenge for teacher educaiton. *Journal of Teacher Education, 60*(5), 497-511.

Ball, D. L., Hill, H. C., & Bass, H. (2005). Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide? *American Educator*.

Ballou, D., & Podgursky, M. (2000). Reforming teacher preparation and licensing: What is the evidence? *Teachers College Record, 102*(1), 5-27.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133-180.

Bennett, C. I., McWhorter, L. M., & Kuykendall, J. A. (2006). Will I ever teach? Latino and African American students' perspectives on PRAXIS I. *American Educational Research Journal, 43*(3), 531-575.

Betebenner, D. W. (2009). *Growth, standards, and accountability*. Dover, NH.

Betebenner, D. W., & Linn, R. L. (2010). *Growth in student achievement: Issues of measurement, longitudinal data analysis, and accountability*. Austin, TX: Center for K-12 Assessment and Performance Management.

Betts, J. R., Zau, A. C., & Rice, L. A. (2003). *Determinants of student achievement: New evidence from San Diego*. San Francisco, CA: Public Policy Institute of California.

Borko, H., Whitcomb, J. A., & Byrnes, K. (2008). Genres of research in teacher education. In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre & K. E. Demers (Eds.), *Handbook of research on teacher education: Enduring question in changing contexts* (3rd ed.). Routledge: New York, NY.

Boyd, D., Goldhaber, D., Lankford, H., & Wyckoff, J. (2007). The effect of certification and preparation on teacher quality. *Future of Children, 17*(1), 45-68.

Boyd, D., Lankford, H., Loeb, S., Rockoff, J., & Wyckoff, J. (2008). The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis and Management, 27*(4), 793-818.

Buddin, R., & Zamarro, G. (2009). Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics, 66*(2), 103-115.

Center for Education Policy and Practice (2008). *Tomorrow's teachers: Preparing the education workforce for 21st Century schools.* Boston, MA: Massachusetts Teachers Association.

Charalambous, C. Y. (2010). Mathematical knowledge for teaching and task unfolding: An exploratory study. *Elementary School Journal, 110*(3), 247-278.

Chester, M. D. (2008). *Celebrating progress, committing to next steps to narrowing achievement gaps.* Paper presented at the Curriculum and Instruction Summit, Marlborough, MA.

Chester, M. D. (2009a). *Massachusetts 2009: The state of education.* Paper presented at the 2009 Curriculum, instruction, and assessment summit, Marlborough, MA.

Chester, M. D. (2009b). Mathematics subtest for elementary and special education teachers: Determination of passing score and proposed amendment to educator licensure regulations for transition period from http://www.doe.mass.edu/boe/docs/FY09/0509.pdf

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

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. [Proceedings Paper]. *Economics of Education Review, 26*(6), 673-682.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Resources, 45*(3), 655-681.

Cochran-Smith, M. (2005). Teacher education and the outcomes trap. *Journal of Teacher Education, 56*(5), 411(417).

Cochran-Smith, M., & Fries, K. (2008). Research on teacher education: Changing times, changing paradigms. In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre & K. E. Demers (Eds.), *Handbook of research on teacher education: Enduring question in changing contexts* (3rd ed.). Routledge: New York, NY.

Cochran-Smith, M., & Zeichner, K. M. (2005). *Studying teacher education: The report of the AERA Panel on Research and Teacher Education*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McParland, J., Mood, A. M., Weinfeld, F. D., et al. (1966). *Equality of Educational Opportunity*. Washingon, D.C.: U.S. Department of Health, Education, and Welfare.

Corcoran, S. P. (2009). Human capital policy and the quality of the teacher workforce. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession*. Washington, DC: The Urban Institutes Press.

Council, N. R. (2001). *Testing teacher candidates: The role of licensure tests in improving teacher quality*. Washington, D.C.: National Academy Press.

Croninger, R. G., Rice, J. K., Rathbun, A., & Nishio, M. (2007). Teacher qualifications and early learning: Effects of certification, degree, and experience on first-grade student achievement. *Economics of Education Review, 26*(3), 312-324.

D'Agostino, J. V., & Powers, S. J. (2009). Predicting teacher performance with test scores and grade point average: A meta analysis. *American Educational Research Journal, 46*(1), 146-182.

Darling-Hammond, L. (1997). *Doing what matters most: Investing in quality teaching*. New York: National Commission on Teaching and America's Future.

Darling-Hammond, L. (2007). We need to invest in math and science teachers. *Chronicle of Higher Education, 54*(17), B20.

Darling-Hammond, L. (2008). Knowledge for teaching: What do we know? In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre & K. E. Demers (Eds.), *Handbook of research on teacher education: Enduring question in changing contexts* (3rd ed.). Routledge: New York, NY.

Darling-Hammond, L., Berry, B., & Thoreson, A. (2001). Does teacher certification matter? Evaluating the evidence. *Educational Evaluation and Policy Analysis, 23*(1), 57-77.

Darling-Hammond, L., Bransford, J. D., Lepage, P., Hammerness, K., & Duffy, H. (Eds.). (2005). *Preparing teachers for a changing world: What teachers should learn and be able to do*. San Francisco: Josey-Bass.

Dee, T. S. (2001). *Teachers, race, and student achievement in a randomized experiment*. Cambridge, MA: National Bureau of Economic Research.

Delpit, L. (2006). Cross-cultural confusions in teacher assessment *Other People's Children: Cultural Conflict in the Classroom* (pp. 135-151). New York: New Press.

Farkas, S., & Duffett, A. (2010). *Cracks in the Ivory Tower? The Views of Education Professors Circa 2010*. Washington, D.C.: Thomas B. Fordham Institute.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

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

Ferguson, R. F. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal on Legislation, 28*(2), 465-498.

Flippo, R. F., & Riccards, M. P. (2000). Initial teacher certification testing in Massachusetts. *Phi Delta Kappan, 82*(1), 34.

Florio-Ruane, S. (2002). More Light: An Argument for Complexity in Studies of Teaching and Teacher Education. *Journal of Teacher Education, 53*(3), 205-215.

Fowler, R. C. (2001). What did the Massachusetts teacher tests say about American education? *Phi Delta Kappan, 82*(10), 773.

Fuhrman, S. H. (2004). Introduction. In S. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 3-14). New York: Teachers College Press.

Goe, L. (2007). *The link between teacher quality and student outcomes : a research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Goldhaber, D. (2004). Why do we license teachers? In F. M. Hess, A. J. Rotherham & K. Walsh (Eds.), *A qualified teacher in every classroom?* Cambridge, MA: Harvard Education Press.

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

146

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. [Article]. *Review of Economics and Statistics, 89*(1), 134-150.

Goldhaber, D., & Hannaway, J. (2009). Overview. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession*. Washington, DC: The Urban Institutes Press.

Goldhaber, D., & Hansen, M. (2009). Race, gender, and teacher testing: How informative a tool is teacher licensure testing? *American educational research journal., 47*(1), 218.

Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources, 32*(3), 505-523.

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129-145.

Goldhaber, D. D., & Brewer, D. J. (2001). Evaluating the evidence on teacher certification: A rejoinder. *Educational Evaluation and Policy Analysis, 23*(1), 79-86.

Goldhaber, D. D., & Hannaway, J. (2009). *Creating a new teaching profession*. Washington, D.C.: Urban Institute Press.

Green, E. (2010, March 2). Building a better teacher. *The New York Times*. Retrieved from http://www.nytimes.com/2010/03/07/magazine/07Teachers-t.html

Greenberg, J., & Walsh, K. (2008). *No common denominator: The preparation of elementary teachers in mathematics by America's education schools*: National Council on Teacher Quality.

Haney, W., Fowler, C., Wheelock, A., Bebell, D., & Malec, N. (1999). Less truth than error: Massachusetts teacher tests. *Educational Policy Analysis Archives, 7*(4).

Hanushek, E. A. (2009). Teacher deselection. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession*. Washington, DC: The Urban Institutes Press.

Hanushek, E. A., & Rivkin, S. G. (2010a). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267-271.

Hanushek, E. A., & Rivkin, S. G. (2010b). The quality and distribution of teachers under the No Child Left Behind Act. *Journal of Economic Perspectives, 24*(3), 133-150.

Harrell, P. E. (2009). Do state examinations measure teacher quality? *Educational Studies, 35*(1), 65-79.

Harrington, K. (1999). The sound and fury of teacher testing in Massachusetts. *Metropolitan Universities: An International Forum, 10*(2), 57-62.

Harris, D. N. (2009). Teacher value-added: Don't end the search before it starts. *Journal of Policy Analysis and Management, 28*(4), 693-711.

Harris, D. N., & Rutledge, S. A. (2010). Models and predictors of teacher effectiveness: A comparison of research about teaching and other occupations. *Teachers College Record, 112*(3), 914-960.

Harris, D. N., & Sass, T. R. (2006). *The effects of teacher training on teacher value-added*. Paper presented at the American Education Finance Association.

Harris, D. N., Sass, T. R. (2007). *Teacher training, teacher quality, and student achievement. Working Paper 3*: National Center for Analysis of Longitudinal Data in Education Research.

Heck, R. H. (2007). Examining the relationship between teacher quality as an organizational property of schools and students' achievement and growth rates. *Educational Administration Quarterly, 43*(4), 399-432.

Herbruck, E. H. (2006). *Performance on Ohio teacher tests: Equity implications for the teacher force*. Unpublished Dissertation, Cleveland State University.

Hess, F. M. (2005). The predictable, but unpredictably personal, politics of teacher licensure. *Journal of Teacher Education, 56*(3), 192.

Hess, F. M., Rotherham, A. J., & Walsh, K. (2004). Introduction. In F. M. Hess, A. J. Rotherham & K. Walsh (Eds.), *A qualified teacher in every classroom?* Cambridge, MA: Harvard Education Press.

Hill, H. C. (2009). Evaluating value-added models: A vaidity argument approach. *Journal of Policy Analysis and Management, 28*(4), 700-709.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., et al. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*(4), 430 - 511.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371-406.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal, 105*, 11-30.

Imig, D. G., & Imig, S. R. (2006). The teacher effectiveness movement - How 80 years of essentialist control have shaped the teacher education profession. *Journal of Teacher Education, 57*(2), 167-180.

Imig, D. G., & Imig, S. R. (2008). From traditional certification to competitive certification: A twenty-five year retrospective. In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre & K. E. Demers (Eds.), *Handbook of research on teacher education: Enduring question in changing contexts* (3rd ed.). Routledge: New York, NY.

Jacob, B. A., & Lefgren, L. (2004). The impact of teacher training on student achievement - Quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources, 39*(1), 50-79.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-136.

Jepsen, C. (2005). Teacher characteristics and student achievement: evidence from teacher surveys. *Journal of Urban Economics, 57*(2), 302-319.

Johnson, S. M., & Kardos, S. M. (2008). The next generation of teachers: Who enters, who stays, and why. In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre & K. E. Demers (Eds.), *Handbook of research on teacher education: Enduring question in changing contexts* (3rd ed.). Routledge: New York, NY.

Kennedy, M. M. (2008). Contributions of qualitative research to research on teacher qualifications. *Educational Evaluation and Policy Analysis, 30*(4), 344-367.

Kennedy, M. M., Ahn, S., & Choi, A. J. (2008). The value added by teacher education. In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre & K. E. Demers (Eds.), *Handbook of research on teacher education: Enduring question in changing contexts* (3rd ed.). Routledge: New York, NY.

Kukla-Acevedo, S. (2009). Do teacher characteristics matter? New results on the effects of teacher preparation on student achievement. *Economics of Education Review, 28*(1), 49-57.

Ladson-Billings, G. (1998). Teaching in dangerous times: Culturally relevant approaches to teacher assessment. *Journal of Negro Education, 67*(3), 255-267.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. [Article]. *Educational Evaluation and Policy Analysis, 24*(1), 37-62.

Luke, D. A. (2004). *Multilevel modeling*: Sage Publications.

Massachusetts Board of Elementary and Secondary Education (2009). *Minutes of the regular meeting of the Massachusetts Board of Elementary and Secondary Education, May 19, 2009*. Retrieved from http://www.doe.mass.edu/boe/docs/default.html?section=archive.

Massachusetts Departmet of Elementary and Secondary Education (2011). *MCAS Student Growth Percentiles: An interpretive guide*. Malden, MA.

Massachusetts Department of Elementary and Secondary Education (2007). DOE to review why minority teacher candidates are struggling with MTEL. from http://www.doe.mass.edu/news/news.aspx?id=3457

Massachusetts Department of Elementary and Secondary Education (2007). *Guidelines for the mathematical preparation of elementary teachers*. Retrieved from http://www.doe.mass.edu/edprep/pd.html.

Measures of Effective Teaching Project (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching Project*.

Melnick, S. L., & Pullin, D. (2000). Can you take dictation? Prescribing teacher quality testing. *Journal of Teacher Education, 51*(4), 262.

Moller, J. (2009). School leadership in an age of accountability: Tensions between managerial and professional accountability. *Journal of Educational Change, 10*(1), 37-46.

Monk, D. H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review, 13*(2), 125-145.

Monk, D. H., & King, J. (1994). Multilevel teacher resource effects on public performance in secondary mathematics and science: The case of teacher subject-matter preparation. In R. Ehrenberg (Ed.), *Contemporary policy issues: Choices and consequences in education* (pp. 29-58). Ithaca, NY: ILR.

Munoz, M., & Chang, F. (2008). The elusive relationship between teacher characteristics and student academic growth: A multilevel growth model for change. *Journal of Personnel Evaluation in Education, 20*, 147-164.

Murname, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. Oxford: Oxford University Press.

Neild, R. C., Farley-Ripple, E. N., & Byrnes, V. (2009). The effect of teacher certification on middle grades achievement in an rrban district. *Educational Policy, 23*(5), 732-760.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. [teacher evaluation, value-added modeling, teacher effectiveness]. *Educational Policy Analysis Archives, 18*(23).

O'Connell, A. A., & McCoach, D. B. (2008). *Multilevel modeling of educational data*. Charlotte, NC: IAP.

O'Day, J., & Smith, M. S. (1993). Systemic reform and educational opportunity. In S. Fuhrman (Ed.), *Designing coherent education policy: Improving the system* (pp. 250-312). San Francisco: Jossey-Bass.

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

Patrick, D., Chester, M. D., & Banta, M. (2010). *Race to the Top phase 2: Application for initial funding*.

Phillips, K. J. R. (2010). What does "Highly Qualified" mean for student achievement? Evaluating the relationships between teacher quality indicators and at-risk students' mathematics and reading achievement gains in first grade. *Elementary School Journal, 110*(4), 464-493.

Podgursky, M. (2005). Teacher licensing in U.S. public schools: The case for simplicity and flexibility. *Peabody Journal of Education, 80*(3), 15-43.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and sata analysis methods* (Second ed.). Thousand Oaks: Sage Publications.

Rice, J. K. (2003). *Teacher quality: Understanding the effectiveness of teacher attributes*. Washington D. C.: Economic Policy Institute.

Rivkin, S. G. (2009). The estimation of teacher value added as a determinant of performance pay. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession*. Washington, DC: The Urban Institutes Press.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247-252.

Rotherham, A. J., & Mead, S. (2004). Back to the future: The history and politics of state teacher licensure and certification. In F. M. Hess, A. J. Rotherham & K. Walsh (Eds.), *A qualified teacher in every classroom?* Cambridge, MA: Harvard Education Press.

Rothstein, J. (2010). Teacher quality in educaitonal production: Tracking, decay, and student achievement. *Quarterly Journal of Economics, 125*(1), 175-214.

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*(8), 1525-1567.

Ruth, M., & Barth, P. (1999) Not good enough: A content analysis of teacher licensing examinations. How teacher licensing tests fall short., *Thinking K-16: Vol. 3*. Washington, DC: Education Trust.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*(3), 299-311.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee value-added assessment system (TVAAS) Database: Implications for Educational Evaluation and Research. *Journal of Personnel Evaluation in Education, 12*(3), 247-256.

Sanders, W. L., Wright, S. P., & Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*(1), 57-67.

Schoen, L., & Fusarelli, L. D. (2008). Innovation, NCLB, and the fear factor: The challenge of leading 21st-century schools in an era of accountability. *Educational Policy, 22*(1), 181-203.

Sedlak, M. W. (2008). Competing visions of purpose, practice, and policy: The history of teacher certification in the United States. In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre & K. E. Demers (Eds.), *Handbook of research on teacher education: Enduring question in changing contexts* (3rd ed.). Routledge: New York, NY.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

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

Skiba, R. J., Simmons, A. B., Ritter, S., Gibb, A. C., Rausch, M. K., Cuadrado, J., et al. (2008). Achieving equity in special education: History, status, and current challenges. *Exceptional Children, 74*(3), 264-288.

Slaughter, R. (2008). *Measuring middle school achievement growth with student growth percentile methodology*. Paper presented at the NERA Conference.

Sleeter, C. (2008). Preparing White teachers for diverse students. In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre & K. E. Demers (Eds.), *Handbook of research on teacher education: Enduring question in changing contexts* (3rd ed.). Routledge: New York, NY.

Stotko, E. M., Ingram, R., & Beaty-O'Ferrall, M. E. (2007). Promising strategies for attracting and retaining successful urban teachers. *Urban Education, 42*(1), 30-51.

Stotsky, S. (2009). Licensure tests for special education teachers: How well they assess knowledge of reading instruction and mathematics. *Journal of Learning Disabilities, 42*(5), 464-474.

Summers, A. A., & Wolfe, B. L. (1975). *Which school resources help learning? Efficiency and equity in Philadelphia public schools*. Philadelphia: Federal Reserve Bank.

Tellez, K. (2003). Three themes on standards in teacher education: Legislative expediency, the role of external review, and test bias in the assessment of pedagogical knowledge. *Teacher Education Quarterly, 30*(1), 9.

Villegas, M. A., & Davis, D. E. (2008). Preparing teachers of color to confront racial/ethnic disparities in educational outcomes. In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre & K. E. Demers (Eds.), *Handbook of research on teacher education: Enduring question in changing contexts* (3rd ed.). Routledge: New York, NY.

Walsh, K. (2004). A candidate-centered model for teacher preparation and licensure. In F. M. Hess, A. J. Rotherham & K. Walsh (Eds.), *A qualified teacher in every classroom?* Cambridge, MA: Harvard Education Press.

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

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*.

Wilson, S. M., Floden, R. E., & Ferrini-Mundy, J. (2002). Teacher preparation research - An insider's view from the outside. *Journal of Teacher Education, 53*(3), 190-204.

Wilson, S. M., & Tamir, E. (2008). The evolving field of teacher education: How understanding challenge(r)s might improve the preparation of teachers. In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre & K. E. Demers (Eds.), *Handbook of research on teacher education: Enduring question in changing contexts* (3rd ed.). Routledge: New York, NY.

Wright, S. P. (2010). *An investigation of two nonparametric regression models for value-added assessment in education*. Cary, NC.

Youngs, P., Odden, A., & Porter, A. C. (2003). State policy related to teacher licensure. *Educational Policy, 17*(2), 217.

Zeichner, K. M. (2003). The adequacies and inadequacies of three current strategies to recruit, prepare, and retain the best teachers for all students. *Teachers College Record, 105*(3), 490-519.

Zumwalt, K., & Craig, E. (2008). Who is teaching? Does it matter? In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre & K. E. Demers (Eds.), *Handbook of research on teacher education: Enduring question in changing contexts* (3rd ed.). Routledge: New York, NY.

Zuzovsky, R., & Libman, Z. (2006). Standards of teaching and teaching tests: Is this the right way to go? *Studies in Educational Evaluation, 32*(1), 37-52.