Spring 2018

Can Schools be Reformed by Reforming Assessment?: The Effects of an Innovative Assessment and Accountability System on Student Achievement Outcomes

Carla Marie Evans
University of New Hampshire, Durham

CAN SCHOOLS BE REFORMED BY REFORMING ASSESSMENT?: THE EFFECTS OF AN INNOVATIVE ASSESSMENT AND ACCOUNTABILITY SYSTEM ON STUDENT ACHIEVEMENT OUTCOMES

BY

CARLA M. EVANS

BS, Gordon College, 2000

MDiv, Gordon-Conwell Theological Seminary, 2003

DISSERTATION

Submitted to the University of New Hampshire

in Partial Fulfillment of

the Requirements for the Degree of

Doctor of Philosophy

in

Education

May, 2018

This dissertation has been examined and approved in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Education by:

Dissertation Co-Chair, Suzanne E. Graham, Associate Professor of Education

Dissertation Co-Chair, Todd A. DeMitchell, Professor of Education

Emilie M. Reagan, Assistant Professor of Education

Hadley J. Solomon, Assistant Professor of Education

Charles DePascale, Senior Associate, National Center for the Improvement of Educational Assessment

On February 1, 2018

Original approval signatures are on file with the University of New Hampshire Graduate School.

DEDICATION

This dissertation is dedicated to my husband, Josh, whose unwavering encouragement, support and sacrifice has made this journey possible. Words cannot express my deep appreciation and respect for you. And to my children—Connor, Drew, and James—may your curiosity and sense of wonder continue and may God use you to promote human flourishing as you grow into men. I also want to dedicate this dissertation to my brother, Ruddy, who has supported this work through many nights of babysitting. I'm proud of your accomplishments and growth in these last years. And, finally, to my mom and dad who are no longer alive, but who would have been so proud—miss you both every day.

ACKNOWLEDGEMENTS

It takes a village to raise a child and it also takes a village to write a dissertation! To my co-chairs, Dr. Suzanne Graham and Dr. Todd DeMitchell, thank you for the hundreds of hours you have spent reading and critiquing my work, offering advice and wise counsel, and encouraging my growth as a scholar. I have learned a lot about the kind of professor I hope to be someday from your examples. This work has greatly benefited from my long conversations with the both of you. Thank you.

To the rest of my dissertation committee—Dr. Charlie DePascale, Dr. Emilie Reagan, and Dr. Hadley Solomon—thank you for your critical feedback on this dissertation, as well as on many other projects. You have each been a source of encouragement during my graduate studies and I will always be extremely grateful for each of you.

To my unofficial mentor, Dr. Scott Marion, thank you for your support on this project and the way you have always treated me as a colleague. You are a good human being.

To my friends and fellow graduate students at the University of New Hampshire, thank you for the laughter and support. In particular, thank you, Dr. Sara Clarke-Vivier, for your drop-in chats and numerous conference sleepovers. I'm proud of you and your accomplishments. And to my office mate, Joy Erickson—thank you for the support and encouragement.

And finally to my lifelong friends and cheerleaders outside of the program—especially Kasey Dillon, Shannon Sturgill, and Charissa Thonus—thank you for your support in these years and decades. Will we look back and savor these moments? The jury is still out on that one, but the jury is not still out on the value of community and friendship. Love you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CBAS	Classroom-based assessment system
CBE	Competency-based education
DOK	Depth of Knowledge
ESSA	Every Student Succeeds Act of 2015
NCLB	No Child Left Behind Act of 2001
NECAP	New England Comprehensive Assessment Program
NHDOE	New Hampshire Department of Education
PACE	Performance Assessment of Competency Education
PBA	Performance-based assessment
SBAC	Smarter Balanced Assessment Consortium
USDOE	United States Department of Education

ABSTRACT

CAN SCHOOLS BE REFORMED BY REFORMING ASSESSMENT?:

EFFECTS OF AN INNOVATIVE ASSESSMENT AND ACCOUNTABILITY SYSTEM ON STUDENT ACHIEVEMENT OUTCOMES

by

Carla M. Evans

University of New Hampshire, May, 2018

The *Every Student Succeeds Act* of 2015 authorizes a pilot program that allows up to seven states to develop innovative assessment and accountability systems. Prior to the official pilot program launch, the U.S. Department of Education approved one pilot program—New Hampshire's Performance Assessment of Competency Education (PACE). To implement the PACE pilot, the New Hampshire Department of Education received a 2-year waiver (2014-2016) from federal statutory requirements related to state annual achievement testing and was granted additional waivers for the 2016-2017 and 2017-2018 school years. The purpose of this study is to investigate the average effect of the PACE pilot on 8th grade student achievement outcomes in mathematics and English language arts during the first two years of implementation. This study also examines the extent to which those average treatment effects vary according to student characteristics and among PACE schools. PACE students are compared to non-PACE students with similar probabilities of being selected into treatment using propensity score methods. Multi-level modeling is then used to estimate the average treatment effect for students receiving either one or two years of treatment. Findings from this study provide preliminary evidence that the PACE pilot is having a positive effect on 8th grade student achievement outcomes in mathematics for some students starting in the second year of implementation and no effect in English language arts. Findings also suggest that students with disabilities that attend PACE schools tend to exhibit

positive differential effects in comparison to students with disabilities in the non-PACE comparison group in both subject areas, although these findings should be considered exploratory due to the small number of PACE IEP students in the sample. Findings also suggest that male students that attend PACE schools tend to exhibit negative differential effects in comparison to female students in the non-PACE comparison group in both subject areas. Results are descriptive not causal, however, findings could be used to provide assurance to key stakeholders that PACE students are provided an equitable opportunity to learn the content standards. Also, because the focus of PACE pilot is on performance assessments used throughout the year, this study provides initial evidence that the learning gains on performance assessments may carry over to the more traditional state standardized tests. Implications for research, policy, and practice are also discussed.

Chapter 1: Introduction

Elementary and secondary schools across the United States face an important call to prepare students for college and careers. Reports about flat-lining or declining achievement in math and reading over time alongside continued achievement gaps prompt efforts to improve student achievement for all students. In response, state and federal policymakers since the 1970s have utilized large-scale assessment in K-12 schools as one policy instrument to leverage instructional change in classrooms (Hamilton, 2003; Supovitz, 2009). As Resnick and Resnick (1992) state: "The power of tests and assessments to influence educators' behavior is precisely what makes them potent tools for educational reform" (p. 56). Performance-based assessments, in particular, have been advanced as one critical element in a "new" paradigm for assessment and accountability that supports meaningful learning and systemic educational change (Darling-Hammond, Wilhoit, & Pittenger, 2014).

Along different lines, but also in response to these challenges, there has been a resurgence in competency-based models of education in schools and related policy contexts (Pace, Moyer, & Williams, 2015; C. Sturgis, 2016; Worthen & Pace, 2014). With roots in the mastery learning movement in the 1970s and 1980s, the more recent competency-based education movement attempts to leverage the efficacy of an individualized approach to education and progression in the curriculum upon demonstration of mastery or proficiency to improve student achievement outcomes for all students (Le, Wolfe, & Steinberg, 2014).

This dissertation operates at the intersection of these policy responses and policy contexts—the design of innovative state assessment and accountability systems. Innovative assessment and accountability systems are also referred to in the literature as "balanced" (Chattergoon & Marion, 2016; Gong, 2010; Stiggins, 2006) or "comprehensive" (Council of Chief State School Officers, 2015). The term "innovative" stems from the recently re-authorized *Every Student Succeeds Act* (2015)

that allows up to seven states to apply for a waiver from federal assessment and accountability regulations in order to pilot innovative systems. Innovative assessment and accountability systems are important because they have the potential to re-align state assessment systems in such a way that there is coherence between the underlying theory of learning, goals and purposes for the assessment system, and design of the assessment system. This allows for positive feedback loops to occur through the curriculum, instruction, and assessment cycle, and for efficiency in the number of assessments required to inform stakeholders about students' progress towards proficiency.

The recurring pattern of policy attention on performance assessments and competency-based education is a phenomenon that Anthony Downs (1972) addresses in his discussion of the five stages of an issue-attention cycle. In the first stage, the problem exists, but the public is either not aware of the problem or the problem does not command concern. Downs (1972) refers to this as the pre-problem stage. Then in stage 2 the public seems to suddenly become aware of the issue in a state of "alarmed discovery and euphoric enthusiasm" (p. 28) with cries for the public problem to be addressed and solved immediately. However, after the public realizes the cost (broadly defined, not just monetary) involved in solving the problem (stage 3), there is a decline in public interest and heightened concern, which paves the way for other issues and problems to command their way into "alarmed discovery and enthusiastic enthusiasm" (stage 4). Finally, in stage 5, the problem that captured the public's interest and attention moves into a "twilight realm of lesser attention or spasmodic recurrences of interest" (Downs, 1972, p. 35). The issue-attention cycle helps to explain why some policies such as state level performance assessment programs and competency-based education are enacted in a heightened state of public concern then seem to fade to the background only to re-emerge again as policy solutions with "euphoric enthusiasm." Innovative assessment and accountability systems are one policy solution that draws on earlier reform movements and is in the early stages of Downs' issue-attention cycle.

This study takes advantage of a natural experiment occurring in New Hampshire to investigate one instantiation of an innovative assessment and accountability system. This innovative assessment and accountability system's theory of action relies on the power of performance-based assessment within a competency-based learning environment to improve student achievement outcomes. In this way, this study makes an important contribution by examining the effects of one type of innovative assessment and accountability system on student learning while it also adds to the research literature about the effects of performance-based assessment and competency-based education on student achievement outcomes.

In this introductory chapter, I provide general background on K-12 educational assessment. I discuss the role of assessment in a learning environment and the historical role of assessment in education. These discussions help to explain the rise of test-based accountability policies as a policy lever to effectuate systemic K-12 school reform. I also briefly discuss how the negative perceived effects of recent test-based accountability policies have led to a new option for states to innovate with regards to their assessment and accountability systems. This background section then leads into the research problem this dissertation addresses. The research problem is situated in the empirical literature on the effects of performance-based assessment programs and competency-based education on K-12 student achievement outcomes. I then describe the purpose and significance of this dissertation alongside a statement of the research questions. I end this introductory chapter by providing an overview of what is included in each of the subsequent chapters.

The Role of Assessment in a Learning Environment

The word assessment is derived from the Latin *assidere*, meaning "to sit beside or with" (Earl, 2003). The image evoked is that of a teacher sitting with his or her students attempting to understand what is happening in the minds of the students. As a result, assessment is a process of reasoning from evidence, is only an estimate of what a person knows and can do, and is imprecise to some degree (National Research Council, 2001). I use the term 'assessment' broadly in accord with the *Standards of Educational and Psychological Testing* to include any "systematic method of obtaining information" or "systematic process to measure or evaluate" student performance, knowledge, and/or ability, for purposes of drawing inferences (AERA, APA, & NCME, 2014, p. 216). Assessment, therefore, includes: teacher created multiple-choice tests, performance-based assessments, observation, and annual standardized achievement tests (to name a few).

Assessment can also serve various purposes related to learning: assessment *of* learning, assessment *for* learning, and assessment *as* learning (Earl, 2003). Assessment *of* learning measures individual student achievement. Its purpose, therefore, is summative, intended to certify learning and report to parents, teachers, administrators, and other stakeholders about students' progress in school (Earl, 2003). Assessment of learning typically takes place at the end of a unit or course and often takes the form of tests or exams with a grade recorded. This type of assessment does not provide a lot of direction for students in terms of how to improve their learning; it simply audits their learning (Baker & Gordon, 2014; Bennet, 2014).

Assessment *for* learning, in contrast, assists learning and shifts assessment from its summative purposes to its formative purposes. Assessment for learning takes place while students are learning, rather than collecting data after students have learned. Therefore, while assessment of learning occurs typically at the end of a unit or course, in assessment for learning, teachers often collect a wide range of data throughout the learning process, rather than at the end, so they can

modify instruction to fit students' needs (Earl, 2003). Formative assessment often provides specific and actionable information that students can use to adjust their learning and address misconceptions. Assessment for learning can also be dynamic in that it doesn't necessarily include only what a student can do on their own, but can also include assessing how a student performs with the assistance of a teacher or peer, or in a group situation.

Assessment *as* learning extends the role of formative assessment by emphasizing the role of the student (Earl, 2003). Specifically, students are positioned as the connector between the assessment and learning process. For example, when students use metacognitive skills to monitor their learning and use feedback to adjust while learning, assessment itself becomes learning. Additionally, student self-assessment allows students to participate in their own assessment, thereby joining in the larger social practice (Lund, 2013).

Sociocultural learning theorists have argued that assessment practices do more than provide information about what a student knows or can do under certain conditions (Lave & Wenger, 1991; Rogoff, 2003; Vygotsky, 1978). Assessment shapes people's understandings about what knowledge is valued, what learning is, and who learners are (Moss, 2008).

Historical Role of Assessment

Historically, selection and certification has probably been the most common role of formal assessment over the years (Gipps, 1999). For example, assessment has played a key role in many countries by controlling access to higher levels of education and professional careers (Gipps, 1999). According to Gipps (1999), examinations were first developed in China around 200 BC to select candidates for government service. The aim was to reduce nepotism resulting from the inequitable distribution of jobs to wealthy candidates or to candidates whose parents were influential. In the United States in the early 1900s, IQ tests were developed and used to sort and track students into college prep or vocational classes.

More recently, however, assessment in schools has been used to control and drive curriculum and instruction (Gipps, 1999). In other words, assessment has more recently been employed as a policy instrument to leverage change in districts, schools, and classrooms (Hamilton, 2003; Supovitz, 2009). For example, beginning in the 1970s, minimum competency testing held students and teachers accountable for performance with tests intended to serve as signals to teachers and students about what should be taught and learned (Hamilton, Stecher, & Klein, 2002). Labeled as measurement-driven instruction in the 1980s, there was widespread belief in the potential for assessments to shape instruction in positive ways (Popham, 1987; Popham, Cruse, Rankin, Sandifer, & Williams, 1985).

In addition to assessment impacting classroom instruction and the curriculum, the rise of test-based accountability policies in the United States from the 1980s to today stem from the belief that assessment can help reform schooling (Linn, 2008). Thus, assessment is not only one aspect of the learning process triangle with curriculum and instruction that shapes what is learned and mastered by students, but the results of assessment can also be used to restructure the allocation of system resources and impact decision-making connected with the system. In this way, assessment supports educational reform through integrated and systemic change to the institution and its practices.

The use of assessment in schools to effectuate school reform is due to the fact that while assessment and accountability are two distinct notions, they are often inseparable in state and federal policy contexts. Assessment "provides a valid set of inferences related to particular expectations for students and schools" (Linn, Baker, & Betebenner, 2002, p. 2). An accountability system is set of policies and practices that are used to "measure and hold schools and school districts responsible for raising the achievement for all students" (The Education Trust, 2016). Accountability systems use assessments to ascertain how well schools are doing and then prescribe what actions must result

from the ratings (Linn, 2008). Standards explain what students should know and be able to do in a particular grade and subject area, assessment measures progress towards those expectations, and accountability assigns the process for responsibility to ensure that the standards are met. Susan Fuhrman and Richard Elmore (2004) note that accountability systems often provide remedies and sanctions for low performance. These sanctions may involve low consequences or high stakes. Consequently, a change or reform in the standards or type of assessments used impacts the system of accountability.

Rise of Test-Based Accountability as a Central Policy Initiative

The reform of K-12 education through accountability has been a consistent theme since the educational call to arms report, *A Nation at Risk* (The National Commission on Excellence in Education, 1983). *A Nation at Risk* argued that the quality of education in the United States was declining rapidly because of low standards, lack of purpose, ineffective use of resources, and a lackluster drive for excellence. The solution promoted was comprehensive school reform to increase student and school performance and maintain America's pre-eminence internationally. While *A Nation at Risk* brought the quality of education (e.g., excellence) into the foreground, the driving force behind the Elementary and Secondary Education Act of 1965, particularly Title I, was equity. Since the 1980s it has been the dual, and sometimes conflicting, goals of excellence and equity that have driven assessment and accountability policy.

One way that comprehensive school reform has been operationalized in education policy is through two inter-related and inter-dependent reforms: standards-based reforms and test-based accountability reforms. The theory of standards-based reform is the alignment between the state content standards, classroom instruction, and assessment to promote both excellence and equity (Chatterji, 2002; Smith & O'Day, 1991). However, it is impossible to observe and evaluate what is taking place every day in every classroom across the United States and so it has been argued that

aligning state achievement tests to the content standards provides an indicator of student academic achievement, school quality, teacher effectiveness, and "is sufficient for demonstrating opportunity to learn" (Hamilton et al., 2002, p. 73). As Joan Herman (2004) states:

> It is only when the content and process of teaching and learning correspond to the standards that students indeed have the *opportunity to learn* what they need in order to be successful. Under these conditions, too, an assessment provides information on how well students are doing relative to the standards and on the extent to which classroom teaching and learning are helping students to attain the standards. All parts of the system are focusing on the same or a similar conception of standards and are in sync with a continuous improvement model. Without such correspondence, the logic of the standards-based system falls apart (italics added; p. 144).

Therefore, one of the possible inferences resulting from student tests scores is that students are provided an equitable opportunity to learn the content standards (Howe, 1994; McDonnell, 1995).

According to Hamilton (2003), test-based accountability can be defined with three components: 1) testing students, 2) publicly reporting school performance, and 3) rewarding or sanctioning individuals or institutions based on some measure of performance or improvement. Test-based accountability is arguably based on the premise that requiring all students to take standardized tests and attaching high-stakes to the results will improve student achievement outcomes (Hamilton, 2003). In this way, testing has become a widely utilized and relatively inexpensive American federal and state policy instrument to leverage change in districts, schools, and classrooms (Supovitz, 2009).

A prime example of how testing has been utilized as a federal policy instrument to effectuate comprehensive school reform is the *No Child Left Behind* (NCLB) Act of 2001. NCLB significantly increased the federal role in holding schools responsible for the academic progress of all students. For example, under NCLB states were required for the first time to test students annually in English language arts (ELA) and math in grades 3-8 and once in high school. States were required to bring all students to the "proficient level" on state tests by the 2013-2014 school year and individual schools had to meet state "adequate yearly progress" (AYP) targets toward this goal. If a school

receiving Title I funds failed to meet their AYP target two years in a row federally mandated sanctions started to roll in (e.g., offer students a choice of other public schools to attend; offer free tutoring to students; state intervention; etc.).

The addition of the school accountability requirements in NCLB represents a change in the existing theory of action in previous re-authorizations of the Elementary and Secondary Education Act of 1965. Requirements for state content standards, state assessments, performance standards and achievement levels were all in place at least once per grade span for math and ELA in the 1994 re-authorization of the Elementary and Secondary Education Act (i.e., *Improving America's School Act*); however, NCLB added the accountability requirements. The theory of action behind NCLB was that through a focus on school achievement, educators and policymakers would improve education. Strategies to accomplish this include: establishing grade level state content standards, requiring large-scale state annual testing in grades 3-8 and once in high school in math and ELA, setting state-level targets for improvement, identifying schools that fail to meet all targets, and implementing school-level rewards and sanctions. And yet, NCLB did not lead to 100% student proficiency by 2013-2014.

Most studies on NCLB indicate that accountability pressure may positively effect student achievement in math to a small degree, but not as frequently in reading (Ladd & Goertz, 2015). Cronin and colleagues (2005) argue this is because math skills are more likely to be learned in the classroom using a well-defined and sequential curriculum approach; whereas, reading skills are more ubiquitous and can be influenced by parent support outside the classroom more easily. In any case, measuring the effects of test-based accountability policies such as NCLB on student achievement outcomes is confounded because multiple reforms were implemented simultaneously. It is difficult, if not impossible, to separate the effects of the testing and accountability system on student learning from other components.

Historically, research has found that high-stakes, large-scale standardized achievement tests have had overwhelmingly negative effects on teaching (Au, 2007; Booher-Jennings, 2005; Diamond & Spillane, 2004; Hamilton et al., 2007; Herman & Golan, 1993; McMurrer, 2007; Pedulla et al., 2003; Shepard & Dougherty, 1991; Stecher & Chun, 2001; Stecher et al., 2008). Some have argued that the negative effects of large-scale assessment on curriculum and instruction occurred because of a fundamental misalignment between the purpose of assessment and the role assessment has played in schools (Resnick & Resnick, 1992; Shepard, 2000). This results in an incoherent system of assessments that are external to the regular teaching and learning cycle, and therefore do not provide useful or timely instructional feedback to teachers, narrow the curriculum to focus on only those standards and subjects tested on state assessments, and drive the teaching and learning of fragmented bits of knowledge rather than deeper learning (Darling-Hammond, Wilhoit, & Pittenger, 2014; Pellegrino, Chudowsky, & Glaser, 2001; Smith & O'Day, 1991). Deeper learning is defined as "the process through which an individual becomes capable of taking what was learned in one situation and applying it to new situations (i.e., transfer)"(National Research Council, 2012, p. 4).

Recently, there has been a backlash from the heavy emphasis on high-stakes achievement tests in the United States. For example, in the so-called "opt-out movement" parents across the nation have "opted" their children out of taking certain standardized tests (Pizmony-Levy & Green Saraisky, 2016). There is a general belief among many educators and parents that there is too much testing in schools (Hart et al., 2015)—so much so that the United Stated Department of Education (USDOE) proposed a 2% cap in 2015 on the percent of instructional time that should be spent on testing (USDOE, 2015).

Assessment and Accountability Systems that Support Meaningful Learning

Darling-Hammond, Wilhoit, and Pittenger (2014) argue that a new approach to accountability for learning should be implemented in state and federal policy contexts. This approach includes accountability for meaningful learning, which requires a better system of assessments that are aligned with higher-order knowledge and skills. Darling-Hammond and colleagues argue this requires that the singular focus on state-level summative achievement tests should be abandoned and authentic performance tasks should be one key design feature of the new system of assessments.

This call for the use of performance assessments in educational reform is not new. There is a long history of educational reformers calling for changes to assessment and accountability systems and for the use of authentic, complex performance assessments (Haertel, 1999; Linn, Baker, & Dunbar, 1991; Resnick & Resnick, 1992). For example, the National Research Council report, *Knowing What Students Know* (2001), argues that new forms of educational assessment and measurement principles need to be constructed and/or utilized in fitting with advances in cognitive (e.g., constructivist and sociocultural learning theory) and measurement sciences. This calls for a paradigm shift with regard to the use and purpose of educational assessment in schools from the dominant 20th century paradigm of social efficiency curriculum, behaviorist learning theory, and scientific measurement to a 21st century paradigm where a reformed vision of curriculum, cognitive and constructivist learning theories, and classroom assessment (including performance assessment) should shape educational assessment (Gipps, 1999; Shepard, 2000).

The newly passed education reform legislation that succeeds NCLB, the *Every Student Succeeds Act* (ESSA) of 2015, addresses calls for a new K-12 assessment and accountability paradigm, advances in cognitive and measurement sciences, concerns about "over-testing," and overreliance on high-stakes state achievement tests by authorizing a pilot program under the Innovative Assessment

and Accountability Demonstration Authority. This pilot program allows up to seven states to apply to implement innovative assessment and accountability systems. These innovative systems may incorporate new measures of student performance that replace annual state-level achievement testing in pilot states, depending upon how the system is designed. For example, Section 1204 of ESSA states that innovative systems may use competency-based or other innovative assessment approaches to make determinations of student proficiency each year. These may include:

> (1) competency-based assessments, instructionally embedded assessments, interim assessments, cumulative year end assessments, or performance-based assessments that combine into an annual summative determination for a student, which may be administered through computer adaptive assessments;
> (2) assessments that validate when students are ready to demonstrate mastery or proficiency and allow for differentiated student support based on individual learning needs.

The rationale provided in earlier draft legislation for the pilot program is that innovative systems allow "the administration of assessments that may measure student mastery of state academic content standards *more effectively* than current state assessments and *better inform* classroom instruction and student supports, ultimately leading to *improved academic outcomes for all students*" (ESSA, 2015, p. 3, emphasis added).

New Hampshire provides an early model of how an innovative assessment and accountability system might be designed under the Innovative Assessment and Accountability Demonstration Authority authorized by Section 1204 of ESSA. In March 2015, the USDOE officially approved New Hampshire's Performance Assessment of Competency Education (PACE) pilot (NHDOE, 2015). The PACE pilot is the first-in-the-nation waiver from federal statutory requirements related to annual state-level achievement testing. The PACE pilot was granted a 2-year waiver for the 2014-2015 and 2015-2016 school years and additional waivers for the 2016-17 and 2017-18 school years.

The PACE pilot has been closely followed in education circles because it offers one strategy for "reducing the nation's reliance on standardized testing while providing assessments that give

meaningful feedback for students, parents, and teachers" (NHDOE, 2015a, p. 1). In the PACE system, locally-designed and curriculum-embedded performance assessments within a competency-based learning environment serve as the cornerstone of the new accountability model (Marion & Leather, 2015). A standardized achievement test is administered once per grade span (elementary, middle school, and high school) and serves as an external audit on the system. In the rest of the grades, determinations of student proficiency in ELA and mathematics are made using local and common assessment data and teacher judgment surveys.

New Hampshire's PACE pilot, and ultimately any state awarded flexibility under the Innovative Assessment and Accountability Demonstration Authority, must demonstrate that all students are exposed to high-quality instruction, have the same opportunity to learn the content standards, and are held to the same performance expectations.

Statement of the Problem

The problem that this dissertation addresses is a lack of research about whether or to what extent innovative assessment and accountability systems that utilize performance-based assessment within a competency-based learning environment provide students with equitable opportunities to learn the content standards. There have been no empirical studies to date on the only instantiation of a current innovative system—NH's PACE pilot. This matters because even though there are concerns about the negative effects of high-stakes standardized tests on teaching and learning, it is illogical to design and implement an innovative assessment and accountability system like NH's PACE pilot if there are negative effects on student learning or for certain subgroups of students over time under an "innovative" system. It is for this reason that a recent formative evaluation of the NH PACE pilot calls for research that externally verifies the impacts of PACE on teaching *and learning* (Becker et al., 2017, p. 33). This dissertation aims to begin that external verification process related to student achievement outcomes.

Over time researchers have collected evidence on the benefits, limitations, and lessons learned from the implementation of state-adopted performance assessment systems. According to Parke and Lane (2008), performance assessments were a major portion of some states' assessment programs prior to the implementation of the *No Child Left Behind* Act in 2002. In 1990, for example, eight states were using some form of performance assessment in math and/or science, and six other states were developing or piloting performance assessments in math, science, reading, and/or writing (Stecher, 2010). At the same time, an additional ten states were exploring the possibility of incorporating performance assessments into their state assessment system (Stecher, 2010). However, the use of large-scale, high-stakes performance assessment has been scaled back over the last twenty-five years, though not eliminated (Stecher, 2010).

There is evidence from prior research on performance assessment programs included in state testing programs from the 1990s that performance assessment programs may have a small positive effect on student achievement in both math and ELA over time. For example, in the two studies that examined the effects on student achievement outcomes directly, one found a small positive effect on one outcome measure in math after one year (d=0.13), but no effect on either outcome measure in reading after one year (Shepard et al., 1995). The other study found a significant increase in average school performance over five years in five subject areas, including: math, reading and writing (Stone & Lane, 2003).

There were many reasons why state-level performance assessment systems were replaced or scaled back in state assessment and accountability systems over the last 15 years despite some early positive effects on student achievement outcomes. These reasons vary state-to-state, but they typically fall into three categories: concerns about technical quality leading to loss of political will, need for individual student proficiency determinations under NCLB, and feasibility of the system at scale.

Concerns about the technical quality of state-level performance assessment systems centered on the reliability and validity of performance assessment scores. For example, Koretz and colleagues examined Vermont's writing and math portfolio program and questioned the reliability of individual ratings (Koretz, Klein, McCaffrey, & Stecher, 1993; Koretz, Stecher, Klein, & McCaffrey, 1994), although Vermont's system was low-stakes and intended for school-level reporting only (see Hill & DePascale, 2003 for a discussion about how school-level results can be reliable even if individual-level results are only modestly reliable). Concerns about the validity of performance assessment scores was also raised in Vermont because differential assistance was offered to students while they completed their portfolios (Stecher & Mitchell, 1995). In fact, 70% of Vermont teachers said they provided differential assistance to students on their portfolios including "scribing, reading, and providing manipulative aids—to help students do their best work" (Stecher & Mitchell, 1995, p. 33). But with assistance comes a potential threat to the validity of score interpretations—whose work is it? (Gearhart & Herman, 1995). Once the technical quality of the assessment system results was questioned, the political will to continue the programs started to unravel.

Another reason why some states moved away from performance assessment systems was due to the regulations under NCLB. For example, Maryland's school performance assessment program used matrix sampling, which does not allow for individual scores to be reported as required in the NCLB legislation. Feasibility was a factor related to the re-design of Vermont's system as the portfolios were philosophically coherent with what educators in Vermont wanted to accomplish, but not practically feasible to sustain over a long period of time (Tung & Stazesky, 2010). Thus, there were many reasons why state performance assessment programs lost momentum in the 1990s.

Most of the research related to competency-based education is from the 1970s and 1980s. Lack of conceptual clarity about defining features of competency-based education and piecemeal implementation affected the efficacy of competence-based education reforms during this time

period (Block, 1978; Spady, 1977, 1978; Spady & Mitchell, 1977). Spady (1978) was pessimistic about the longevity of competency-based education reforms not because he didn't believe it could transform the educational system, but precisely because it would require "educators and the public to give up decades of habits and assumptions regarding the structures and methods of schooling, just at the time when accountability looks cheaper and safer than another version of school reform" (p. 22). And this is what happened when *A Nation at Risk* was released in 1983, many policymakers turned to accountability as the answer to education reform and away from other education reforms, including competency-based education.

Most of the research on competency-based education as defined and implemented in the 1970s and 1980s comes from the mastery learning movement. This movement focused on the element of time and the need to restructure the school system so that mastery of content was the emphasis not how many school days a student completed. Multiple research reviews found that there were positive effects from mastery-based curricula, including stronger effects for low-achieving students (Anderson, 1994; Block & Burns, 1976; Cotton & Savard, 1982; Guskey & Gates, 1986; Guskey & Pigott, 1988; Kulik, Kulik, & Bangert-Drowns, 1990; Slavin, 1987a). However, the effects of mastery learning on student achievement varied as a function of the type of outcome measure used in the study. Most studies on the effects of mastery learning used locally constructed outcome measures. A few used both locally constructed and standardized achievement tests. Slavin (1987a) was the first to examine effects based upon the type of outcome measure used. Slavin found that effects of mastery learning were small, but positive on locally constructed exams, but effects were trivial on standardized achievement tests and not significant.

Within the last ten years, conversations at the national, state, and local level about competency-based education have re-emerged (Bramante & Colby, 2012; Pace & Worthen, 2014; Chris Sturgis, 2016). These conversations are oftentimes framed within an equity argument whereby

achievement gaps along socioeconomic, ethnic/racial, disability, and English proficiency lines can be addressed when the traditional time-based structure of the American school system is replaced with a competency-based, mastery-based, or proficiency-based approach to education where a personalized approach to education can address individual student needs (Lewis et al., 2014).

There is not a lot of research to-date on the more recent instantiation of competency-based education except for three separate studies that recently examined student achievement outcomes associated with competency-based education reforms (Bill & Melinda Gates Foundation, 2014; Haystead, 2010; Pane, Steiner, Baired, & Hamilton, 2015; Steele et al., 2014). Findings from these studies were generally inconclusive. However, there is some evidence to suggest that there might be small positive effects of competency-based education on K-12 student achievement in reading and math in charter schools founded with competency-based education models after two years of implementation (Bill & Melinda Gates Foundation, 2014; Pane et al., 2015; Steele et al., 2014-Denver & Houston). There is not enough evidence yet to speculate about effects of competency-based education models on K-12 student achievement outcomes in public schools not founded as competency-based or personalized learning schools. Similar to the research on mastery learning, there is some evidence to suggest that effects may be greater for elementary students than middle school and high school students and that the lowest performing students may benefit the most (Bill & Melinda Gates Foundation, 2014; Pane et al., 2015). However, the level and scope of implementation in the three recent studies is a concern in making any generalizations from the research.

The limitations of the prior research literature on state level performance assessment programs and competency-based education are both substantive and methodological. Regarding substantive limitations, there are key differences between past state level performance assessment and competency-based education reforms and current reform efforts in these two areas. Moreover,

the combination of these two reforms in one innovative system has not been examined in the prior research literature. This substantive gap in the literature supports the need for further research in this area.

The methodological limitations of both bodies of prior literature fall into three main categories: 1) lack of appropriate comparisons between treatment and comparison groups that lead to potentially biased treatment effects; 2) lack of student-level analyses, as well as examination of dosage effects and non-linear treatment effects; and 3) lack of consideration of differential effects for students according to disability status, gender, free- and reduced-price lunch status, and prior achievement. Based upon findings briefly summarized above and the limitations of studies to date, there is a need for further research on the effects of competency-based education on student achievement outcomes in all grade levels and all subject areas.

Purpose and Significance of the Study

The purpose of this dissertation, therefore, is to investigate the effects of an innovative assessment and accountability system on student achievement outcomes in math and English language arts. Specifically, this dissertation examines the extent to which structuring an innovative assessment and accountability system around performance-based assessments and competency-based education affects academic achievement and learning outcomes for students. In this dissertation, students attending New Hampshire's PACE pilot schools are considered the treatment students and the comparison students are students with similar probabilities of being selected into treatment but who attend non-PACE schools in New Hampshire. The outcome variables are measurable factors directly related to the purpose for implementing the accountability system in the first place—improved student achievement outcomes.

This study adds to the research base on the effects of state-level performance assessment programs and competency-based education by describing the outcomes from an innovative pilot

program carried out in select New Hampshire school districts during the 2014-15 and 2015-16 school years. This study provides a descriptive (non-causal) examination of student outcomes following exposure to the treatment in the first two years of the pilot program.

The NH PACE pilot is closely watched by educators nationwide as a potential model of an innovative assessment and accountability system that utilizes locally-designed and curriculum-embedded performance-based assessments to produce annual determinations of student proficiency (Rothman & Marion, 2016). While new systems of assessments have great potential to minimize the negative side effects of state annual achievement tests and maximize the instructional usefulness, quality, and timeliness of assessment for accountability purposes, the effects of performance-based assessments on student achievement outcomes in an accountability context has not been explored since the early 1990s—a very different policy context. Moreover, the effects of competency-based education are unclear from the prior literature. This research may provide the empirical evidence that other states need to move forward with plans to develop innovative assessment and accountability models under the Demonstration Authority of the *Every Student Succeeds Act*. The findings may also provide assurance to the U.S. Department of Education that the use of local assessment data for accountability purposes provides all students with an equitable opportunity to learn the content standards and does not harm subgroups of students who are generally considered more at risk in terms of educational disparities.

Research Questions

In order to investigate the effects of the PACE pilot on student achievement outcomes, this dissertation study focuses on three research questions:

- **Research Question #1**: What is the average treatment effect of the PACE pilot on Grade 8 student achievement in mathematics and English language arts when comparing PACE students to non-PACE comparison students with similar probabilities of being selected into treatment?
- **Research Question #2:** Does the average treatment effect vary based on student-level characteristics such as prior achievement, gender, socioeconomic status, race/ethnicity, or disability status?
- **Research Question #3**: How do average treatment effects vary among PACE schools?

This dissertation focuses on 8th grade students during the 2014-15 and 2015-16 school years because of the way this specific innovative assessment system is designed. Students in New Hampshire's PACE pilot only take a state-level achievement test once per grade span: 3rd grade ELA, 4th grade math, 8th grade ELA and math, and 11th grade ELA and math. Students in 3rd and 4th grade have no prior achievement test scores and 11th grade students take SATs, which is not specifically aligned to the Common Core State Standards, but is intended to predict college success (Shaw, 2015). Eighth grade, on the other hand, includes both ELA and math and there are prior achievement test scores. It is for this reason why this study was de-limited to 8th grade students over the first two years of the PACE pilot. Data from the 2016-17 school year was not included because it was not available at the time of this study.

Dissertation Overview

In this chapter, I provided general background on the theoretical and historical roles and purposes of educational assessment in K-12 schools in the United States. I explained how test-based accountability policies have attempted to effectuate systemic school reform, but have created the conditions whereby some educational researchers and policymakers call for a new system of assessments and accountability policies. I outlined the problem and rationale for examining the effects of newly authorized innovative assessment and accountability systems. I also briefly reviewed the prior literature, purpose and significance of this study, and research questions.

In Chapter Two, I review the empirical literature on the effects of performance assessment programs and competency-based education on student achievement outcomes. I pay particular attention to the study designs and methodologies, drawing implications for this study design. Chapter Two ends with a synthesis across the two main bodies of literature related to this dissertation. I note what is understood and what is yet to understand and how that provides a rationale for this dissertation's design.

In Chapter Three, I present the study context, datasets, population, sample, measures, and analytic approach used in this study. I provide a step-by-step description of the analytic methods employed in this dissertation and how those methods address the research questions. I also explain how the analytic sample was carefully identified to create roughly equivalent treatment and comparison groups at baseline. Because the pre-existing differences between the two groups are not equivalent at baseline, this study is descriptive and should not be interpreted as making any causal claims.

In Chapter Four, I address the three research questions. Findings are presented within Chapter Four by subject area (math first and then ELA). To address the first research question, I investigate the average effect of New Hampshire's PACE pilot on student achievement in math and

ELA after the first two years (2014-15 and 2015-16 school years). I compare PACE student performance with non-PACE student performance for students with similar probabilities of being selected into treatment. This provides insight into whether the PACE pilot is having its intended effect on student achievement, on average. To address the second research question, I explore variability in average effects according to student characteristics such as disability status, gender, free- and reduced-price lunch, and prior achievement. This provides insight into how achievement gaps for certain subgroups of students may be narrowing, widening, or remaining constant for PACE students in comparison to non-PACE students. To address the third research question, I examine differences between predicted and observed school-level performance in math and ELA among PACE schools. This provides insight into the extent to which PACE schools perform better or worse than predicted and if there are any trends or patterns in the first two years.

Overall, findings suggest that PACE students tend to perform lower than their non-PACE comparison peers in Year 1 of the pilot. This most likely reflects that students received only one month of PACE treatment during that school year rather than an implementation dip since the PACE pilot was not officially approved until March 2015—about a month before students took the standardized outcome measure. Findings also suggest that starting in Year 2, there are small positive effects of PACE in Grade 8 math for the average student (d=0.14), but basically no effect in Grade 8 English language arts. Results also point to positive differential effects for students with disabilities in Grade 8 math (d=0.20 to 0.50) and Grade 8 ELA (d=0.09 to 0.16), but negative effects for male students that off-set positive treatment effects in Year 2. The findings for students with disabilities should be considered exploratory and in need of replication due to the small number of PACE IEP students in the sample. There are mixed and inconclusive findings based on the other student-level characteristics examined—prior achievement and free- and reduced-price lunch. For schools implementing PACE in both years of the pilot, there is some evidence to suggest that schools

perform better than expected starting in the second year of implementation, although the sample size is limited and findings are not generalizable.

In Chapter Five, I conclude that results could provide assurance to key stakeholders that PACE students are provided an equitable opportunity to learn the content standards. I also conclude that these results provide initial evidence that the learning gains exhibited by students because of a performance assessment program and/or competency-based learning environment may be transferring over to the state annual achievement test. I discuss limitations of this study, as well as implications for research, policy, and practice. Specifically, I argue that more research needs to be conducted over time and in other grades and subject areas to examine whether positive effects accumulate over time and the extent to which school-level achievement trends continue to grow based on years of implementation.

Chapter 2: Literature Review

In Chapter One, the research problem and research questions were presented. In this chapter, the relevant empirical literature is presented and critiqued. This dissertation investigates the effects of an innovative assessment and accountability system on student achievement outcomes and if those effects vary according to observable student characteristics. There are many possible bodies of literature that pertain to innovative assessment and accountability systems. For example, there is a large body of literature on prior innovative assessment and accountability systems during the 1990s (e.g., Borko & Elliott, 1998; Borko, Elliott, & Uchiyama, 2002; Firestone, Mayrowetz, & Fairman, 1998; Koretz, Barron, Mitchell, & Stecher, 1996; Koretz, Stecher, Klein, & McCaffrey, 1994; Koretz, Stecher, Klein, Mccaffrey, & Deibert, 1993; Smith et al., 1997; Stecher & Mitchell, 1995). There is also literature related to the technical concerns of those prior systems such as the reliable scoring of performance-based assessments used in an accountability context (e.g., Davey et al., 2015; Hambleton et al., 1995; Koretz, Klein, McCaffrey, & Stecher, 1993; Koretz, McCaffrey, Klein, Bell, & Stecher, 1992; Stecher, 2010).

However, this dissertation investigates the effects of an innovative assessment and accountability system designed around performance-based assessments and competency-based education on K-12 student achievement outcomes in English language arts and mathematics. As a result, this literature review focuses on these two main bodies of literature: effects of performance assessment programs and K-12 competency-based education on student achievement outcomes. The following five inclusion and exclusion criteria were used to cull the literature:

(1) The study takes place in the United States. Studies from outside the United States were not included in this literature review because the socio-political contexts are different.
(2) The study examined either a performance assessment program at the school-, district, or state-level or examined competency-based education learning environments. Further explanation of

these inclusion/exclusion criteria specific to each area of research is delineated at the beginning of each section below.

(3) The study investigates the effects (not perceived effects) of performance assessment programs or competency-based education learning environments on K-12 student achievement outcomes in English language arts or mathematics. This means student test scores had to be included as the main outcome measure and English language arts or mathematics had to be examined.

(4) The study uses quantitative methodology to examine the effects on K-12 student achievement outcomes. Studies that use only qualitative methodology are excluded (e.g., Khattri, Kane, & Reeve, 1995) as are studies that focus on postsecondary outcomes (e.g., Kulik et al., 1979).

(5) The study was published in a peer-reviewed journal article, non-published dissertation, non-refereed research report, or book. No time span was delimited.

In terms of the search methods, the search terms "performance based assessment" or "performance assessment" or "competency based education" or "mastery learning" and "academic achievement" and "elementary secondary education" were used in scholarly databases including ERIC, JSTOR, PsycINFO, ProQuest's dissertation abstracts, and Google Scholar. The same search terms were used in the Google search bar to identify relevant research reports not identified through scholarly databases. Also, the reference lists for each relevant publication were used to identify other possible sources.

Once sources were identified, the titles and abstracts were examined for relevance and adherence to the search inclusion/exclusion criteria. Over 100 abstracts were reviewed, but ultimately only a small number of studies met the inclusion/exclusion criteria. Overall, I reviewed seven research studies related to performance assessment programs, seven research reviews on mastery learning programs, and three research studies on competency-based education programs.

The small number of research studies in these areas foreshadows the need for additional empirical research in these areas.

Organization of the Literature Review

In the first section of the literature review, I discuss the history of and rationale for performance-based assessment. This discussion sets the context for a detailed and thorough review of the quantitative research on the effects of performance assessment programs on student achievement outcomes. This review is organized by state and then chronologically to emphasize how the research is embedded in the larger socio-political context and follows a prescribed research trajectory. Study purpose, research question(s), data sources, methods, findings, and research design strengths and limitations are highlighted. The first section ends with a synthesis across the research in this area with an emphasis on what is known and what is left to understand.

The second section begins with a brief overview of the history and resurgence of competency-based education in the United States. As in this first section, this discussion then sets the context for a detailed and thorough review of the empirical research on the effects of competency-based education on student achievement outcomes. This review is organized chronologically to emphasize the emergence of a new strand of competency-based education studies in the 2010s that build upon mastery learning studies in the 1970s and 1980s. The same process detailed above is followed, including: explaining study purpose, research questions, data sources, methods, and findings, as well as highlighting research design strengths and limitations. This section also ends with a synthesis across the research in this area to detail what is known and what is left to understand.

The third section synthesizes across the two bodies of literature to draw out implications of the prior literature for this dissertation. This section explains how the research methods in this

dissertation build on and address the strengths and limitations of prior research, as well as what the prior literature foreshadows in terms of the expected findings from this dissertation.

Background on Performance-Based Assessment in State Assessment Programs

In the early 1980s, performance-based assessments were thought to be a very promising alternative to standardized tests based primarily at first on evidence of their construct validity and then later because of their potential to influence teaching and learning (Herman, 2004). Performance-based assessments are typically multi-step tasks that require students to produce a product or carry out a complex performance as a demonstration that the instructional goal has been learned (Stecher, 2010). Examples include open-ended problems, essays, and hands-on science experiments (to name a few). They are typically scored through teacher (or rater) judgment using pre-specified criteria, often in the form of a scoring guide or rubric, although computer-automated scoring procedures have been used to reduce the costs associated with scoring (Lane & Stone, 2006). Some performance-based assessments require extended time to complete the task while others are relatively short in duration.

Performance-based assessments are considered "authentic" because it is assumed that the act of completing the assessment is a worthwhile task in and of itself; in other words, the performance that is observed is closely related to the performance of interest (Resnick & Resnick, 1992; Wiggins, 1992). Performance assessment then is thought to be a more *direct* measure of student performance rather than just an indicator of performance as is the case with a standardized achievement test (Lane & Stone, 2006). For this reason, performance assessment has been highly valued for measuring complex performance in the educational measurement community for a long time (Linn et al., 1991).

As mentioned in Chapter 1, performance-based assessments were a major portion of some states' assessment programs prior to the implementation of the NCLB (Parke & Lane, 2008; Stecher,

2010). In 1990, for example, eight states were using some form of performance assessment in English language arts, math and/or science, and six other states were developing or piloting performance tasks in math, science, reading, and/or writing (Stecher, 2010). At the same time, an additional ten states explored the possibility of incorporating performance assessments into their state assessment system (Stecher, 2010). However, the use of large-scale, high-stakes performance-based assessments has been scaled back over the last fifteen years, although not eliminated (Stecher, 2010). Some have argued that NCLB was a factor in state decisions to significantly reduce performance assessment programs (Parke, Lane, & Stone, 2006; Rothman & Marion, 2016; Stecher, 2010). For example, NCLB required all students in grades 3 to 8 and once in high school to have individual scores in reading/writing and math, but states like Maryland used matrix sampling in their performance assessment program and only reported scores at the school level (Stecher, 2010). Additional concerns about the technical quality and cost of performance-based assessments, resources for professional development at scale, as well as swings in political leadership and resolve affected the use of performance assessments in state assessment and accountability systems (Tung & Stazesky, 2010).

Review of the Research Literature on Effects of Performance Assessment Programs on Student Achievement

The main delimiting criterion for inclusion in this review of the performance assessment program research literature was that the study had to investigate the *effects* of the performance assessment program on K-12 student achievement, which means student test scores had to be included as the main outcome measure. Performance assessment programs are defined as the systematic use of performance-based assessments for summative accountability purposes at the school-, district-, or state-level. The accountability context could be either high-stakes (e.g., school-level accountability) or low-stakes (e.g., providing comparative information about the relative performance of schools and districts, but without consequences).

Most of the research in this field of study takes place within the states that experimented with some form of performance assessment in their large-scale testing program starting in the 1990s (i.e., California, Connecticut, Kentucky, Maryland, North Carolina, Vermont, Washington state). However, studies in only three of those states (Kentucky, Washington State, and Maryland) investigated the *effects* of those performance assessment programs on K-12 student achievement outcomes using student test scores as the main outcome measure. This review is organized by those three states starting with the earliest implementer (Kentucky) and then chronologically within each state to emphasize how the research is embedded in the larger socio-political context and follows a prescribed research trajectory. I describe each state-adopted performance assessment program as it arises to provide a holistic view of the program that can then serve as a contextual foundation for comparison of study findings. The one exception is the first study that I will review, which is a one-year intervention design in Colorado.

Review of Shepard et al. (1995) Study

The first empirical research I located examined student achievement resulting from a performance-based assessment program was published in 1995. This was a school-level intervention. In this mixed-methods study, Shepard and colleagues (1995) used a one-year intervention design during the 1992-1993 school year to examine the claim that authentic assessment improves instruction and student learning. They argued that the research literature to date had only inferred the benefits of performance assessments by analogy from research documenting negative effects of traditional, multiple choice tests. The researchers stated that there were no empirical studies on the relationships between performance assessments and student learning to date. The researchers also adopted the perspective that it is not the high-stakes accountability pressure associated with performance assessments that leverages changes in student learning, but "the informational and feedback effects of classroom-embedded assessments" (p. 3).

There were five interrelated research questions in the Shepard et al. (1995) study. This review focuses on the question, "Did students learn more because performance assessments were used in classrooms?" In order to investigate this question, the researchers selected a school district in Colorado that "was known for its extensive mastery learning and criterion-referenced testing system" in the 1980s (p. 4). Mastery learning is the precursor to competency education. A requirement of district participation was the district's willingness to apply to the state for a 2-year waiver from standardized testing in the schools that chose to participate so that teachers could operate in a low-stakes context. Third-grade was chosen because district standardized testing occurred only once per grade span and the researchers wanted to utilize multiple outcome measures, including district testing results. Schools could apply to participate if every third-grade teacher was willing to commit to the intervention. The assessment project's one-year intervention included weekly professional development for teachers on how to create and use performance assessments as part of their instruction. The intervention did not provide a pre-packaged curriculum and assessment package, or focus on changing the curriculum or other instructional practices.

The sample included 13 third-grade classrooms that self-selected into the project from 3 schools in the Colorado district (N=335 third-grade students). The researchers compared the gains in student achievement from the participating schools to gains in student achievement from "matched" control schools. Matches were not exact, but the closest they could find. Schools were matched only on socioeconomic factors (percent free and reduced lunch) and percent minority because it was impossible to match schools on multiple dimensions. For example, school average prior academic achievement could not be used as a matching category because the schools were too different.

Shepard and colleagues (1995) found that after one year there was no difference in student learning in reading or mathematics from the performance assessment intervention indicated on *both*

outcome measures. However, they did find a small effect on math achievement on one outcome measure (d=0.13), but no effect in math on the alternative math test. There were no effects for reading on either outcome measure. This suggests that it may be more likely to see gains in math after one year than in reading. The researchers argued that the "small year-to-year gain in mathematics...helped participating students catch up to the control students in math achievement" (p. 12). The researchers also examined the effects by classroom and found mixed effect sizes in the participating math classrooms—half gained a great deal (d=0.25 to 0.50), but the other half gained zero or lost ground. This suggests that gains may vary according to classroom-based factors that may not be accounted for in the study design such as fidelity-of-implementation.

The researchers conclude: "It is clear that introducing performance measures did not produce immediate and automatic improvements in student learning. This finding should be sobering for advocates who look to changes in assessment as the primary lever for educational reform" (p. 15). That said, they do point out one mitigating factor that they "did not teach to the project outcome measures" and the results were not going to be reported in the context of school accountability—a high-stakes use of project results. Additionally, there were also no curricular or instructional changes promoted alongside the "intervention," which may have weakened the project's effect.

Limitations and Implications of the Shepard et al. (1995) Study

Shepard et al.'s study (1995) supports the design of my study for at least two reasons. First, because they argued for the use of matched-controls, although this study was conducted prior to the widespread use of propensity score methods. Second, the researchers argued for the importance of an outcome measure that is sensitive to pick up the effects of a performance assessment "intervention" project while also realistically measuring the learning in control schools.

One methodological advancement of my study is the identification of the comparison group. In the Shepard et al. (1995) study, the researchers identify "matched" controls at the school-level, but their matching is not exact. This means comparisons between the treatment/intervention group and the comparison group may result from pre-existing differences between the two groups rather than the intervention's efficacy. For example, participating School 1 had 61% of students qualify for free and reduced lunch while the control School 1 had 55% of student qualify for free and reduced lunch. Also, their matching was limited in terms of dimensionality because the limited number of schools to choose from in one district made it impossible to find exact matches on many dimensions at once. For example, the researchers were unable to match schools based on prior achievement, which other research suggests may have a sizable impact on student achievement if it is below average (Allensworth, Moore, Sartain, & Torre, 2016). This study extends and improves upon the Shepard et al. (1995) study because of the use of propensity score methods that attempts to create equivalent treatment and comparison groups at baseline based on many dimensions at once (prior achievement, free- and reduced-lunch status, individualized education plan status, limited English proficiency status, race/ethnicity, gender, etc.) so that unbiased estimates of treatment effects can be made.

Also, because Shepard et al. (1995) found almost no effects after one year, a longer period of time to study program effects is warranted. It may be the case that the use of performance assessments does not provide immediate results and, like many interventions, it takes time for the reform to percolate and for effects to occur (if they are going to occur). This study builds upon the Shepard et al. study by tracking program effects over two years. The cohort implementation strategy whereby groups of districts begin implementing in different years, allows dosage effects to be tested to ascertain if more years in the project amount to different levels of program effects.

In addition, in the Shepard et al 1995 study, the project intervention did not attempt to influence curriculum or other instructional practices just the use of performance assessments. It is unclear, therefore, what effect a reform might have that also aims to change curriculum and instruction. Furthermore, the Shepard study did not occur in the context of a high-stakes school accountability context so it is unclear if teachers would be more motivated to change their curricular, instructional, and/or assessment practices in a different accountability setting such as the one investigated in this dissertation.

One justification for the use of Smarter Balanced (SBAC) achievement tests as an appropriate and valid outcome measure to estimate project effects in this dissertation is that it is reasonable to assume that teachers in both PACE and non-PACE schools would be equally motivated to "teach to the test" since SBAC is used to produce annual determinations of student proficiency (a high-stakes accountability purpose); whereas, in the Shepard et al 1995 study, the use of an independent, alternate non-accountability measure that teachers could not "teach to" was described as a potential mitigating factor limiting program effects. Although there are well documented concerns with "teaching to a test" and how doing so may cloud the validity of test score gains (Jennings & Bearak, 2014; Koretz, 2005), if SBAC measures the breadth and depth of the content standards and well-represents the knowledge and content domain, then increasing scores on SBAC would validly reflect improvements in students' understanding and therefore serves as an appropriate and valid outcome measure of program effects.

Kentucky Instructional Results Information System (KIRIS)

The Kentucky Reform Act of 1990 emerged from a 1989 decision by the state's Supreme Court that declared the education system was unconstitutional (Stecher, 2010). The Kentucky Instructional Results Information System (KIRIS) was mandated from the Kentucky Reform Act and was implemented from 1992-1999. KIRIS tested students in writing in grades 4, 7, and 11 and

math in 5, 8, and 11 using mainly a three-part assessment (Stecher, 2010). One part of KIRIS was a state standardized, on-demand test with multiple-choice and constructed-response items. The other two parts of KIRIS were performance assessments. There were writing and math portfolios for students in the grades specified above that were holistically scored at the local level with one single score reported (Tung & Stazesky, 2010). Students included six pieces/types of writing in the portfolio and specific guidelines were given (Stecher, 2010). There were also short and extended performance tasks that were centrally scored in math and English language arts (Tung & Stazesky, 2010). The extended performance tasks were administered eight times per year in the tested grades and subjects and reported only at the school level. Students worked both collaboratively and individually on the tasks, which changed each year (Tung & Stazesky, 2010).

Kentucky partnered with universities and non-profits to provide professional development to teachers implementing KIRIS (Borko et al., 2002). One unique feature of the professional development model was that 65% of the state funds were directed to schools to spend as they deemed best (Borko et al., 2002). Stakes were attached to the results from KIRIS at the school-level with rewards and sanctions for schools that did or did not meet performance expectations respectively. KIRIS lost political support from policymakers and parents as a result of concerns over the technical quality of the system (Stecher, 2010). KIRIS was replaced in 1999 with another assessment and accountability system that kept some of the components, but eliminated the math portfolios. That system was later replaced with a test for NCLB reporting that was mainly multiple-choice with some constructed response items (Stecher, 2010).

Review of Stecher et al. (1998) Kentucky Study

In 1995, Researchers at RAND and the University of Colorado Boulder, under the auspices of CRESST (Center for Research on Evaluation, Standards, and Student Testing) started researching new standards-based assessment and accountability systems. The researchers studied reforms

sequentially in two states using similar investigations and survey instruments: Kentucky, an early implementer of standard-based reform, from 1995 to 1998 with Washington state research beginning in 1998. Similar to the study context of this dissertation, both states adopted performance- based assessments in order to drive changes in instruction.

In order to investigate which classroom practices (standards-based vs. traditional) were associated with improvements in assessment results (KIRIS gains), Stecher and colleagues (1998) used both teacher surveys and KIRIS accountability index gains. They were interested in exploring the effects of particular reform-oriented practices (including using performance-based assessments) on student achievement at the school-level. To do so, the researchers surveyed a representative sample of about 560 teachers from across the state of Kentucky during the 1996-97 school year. Surveys were sent to elementary and middle schoolteachers in the KIRIS accountability grades and subject areas: writing (4th and 7th grade) and math (5th and 8th grade). Two stratification variables were used to draw the sample: gain on the KIRIS accountability index in the subject of interest and school size. Schools were placed in three equal strata (low, medium, and high) based on their gain in writing or math during the second biennial (or every other year) accountability cycle (1992-94 vs. 1994-96). Schools were also placed into two equal strata (small and large) based on school size. Within each stratum a random sample of schools were chosen. For each of the survey populations (four grade/ subject combinations), approximately 70 schools were selected. No school was selected for more than one sample. Low- and high-gain schools were over-sampled to increase the power for detecting differences in classroom practices between low- and high-gain schools, which was the focus of the research questions. Overall, there was about a 70% teacher response rate (RR) with about 400 teachers responding.

Cases were weighted prior to analysis because of the intentional over-sampling of high- and low-gain schools. Researchers calculated descriptive statistics separately for each grade and for

teachers in high- and low-gain schools within each grade. The researchers also combined grades to compute statistics by subject. The analysis tested the significance of differences between responses for teachers in high- and low-gain schools based on second biennium gain scores using chi-square tests and t-tests, as appropriate. For example, differences between mean scores on high- and low-gain schools were tested using *t*-tests. It appears that the researchers used individual survey items to measure classroom practices in this study rather than a composite of items. In the next study these researchers conducted in Kentucky the following year, they created composite measures using similar survey items of classroom practices and a different analytic approach. No information on the reliability or construct validity of the survey instrument was provided.

Stecher and colleagues (1998) found no consistent differences between classroom practices in high- vs. low-gain schools based on KIRIS gains. For example, there were cases where standards-based practices were associated with high gains in one subject or grade level and some cases where traditional practices were associated with high gains. There were also some associations that were counter-intuitive. For example, teachers in low-gain schools were more positive about the impact of writing portfolios than teachers in high-gain schools. Overall, the researchers state: "We did not find convincing evidence that a particular set of actions or policies would produce higher scores. If there is such a pattern it would appear to include both standards-based and traditional approaches" (p. 85). These findings may be an artifact of the survey instrument itself, however, especially as no evidence is provided of the survey's internal consistency (reliability) and construct validity.

The researchers also go on to specify a few reasons why they may have failed to detect relationships that are really there, including "the volatility of gain scores, the sensitivity of our instruments, and the timing of our survey" (Stecher et al., 1998, p. 85). The researchers explain this statement in that they compared self-reported practices in 1996-1997 with school-level gain scores from 1992-1994 to 1994-1996. Because of the biennial (every two year) implementation of KIRIS

this means they had only two time points. As in the Shepard et al. (1995) study, this suggests that baseline and one time point, or one year of data, may be too little to examine program effects.

Limitations and Implications of the Stecher et al. (1998) Kentucky Study

There are several limitations of the Stecher et al. (1998) Kentucky study. First, the researchers relied on principals to provide teachers' names, however, not all principals provided teacher names and not all teachers who were surveyed chose to participate. Therefore, selection bias may threaten the internal validity of study findings. Also, there is more than one teacher surveyed in many schools, which may result in a small clustering effect[1] that is not accounted for in the study design, but may bias study findings. Self-report data is also particularly perceptible to social desirability bias and memory effects. The researchers also created a categorical variable (high- vs. low-gain schools) from continuous data, which limits how much variability can be explained or predicted by subsequent analyses.

The ways in which my study extends and improves upon the Stecher et al. (1998) study, however, is complicated by major differences in our study's purposes and research questions. One of the major differences between the Stecher et al. (1998) study and my dissertation study is that Stecher and colleagues did not compare treatment versus comparison schools. There were no comparison schools; all schools in Kentucky were state-mandated to adopt KIRIS as it was not voluntary. Therefore, Stecher and colleagues could not examine the average effect of KIRIS in treatment versus comparison schools. Instead, they examined if there were any differences in how high- vs. low-gain schools reported use of standards-based and traditional-based classroom practices. They found there was no clear pattern.

Given differences in study purposes, one methodological improvement of this study is I do not create categorical variables from continuous data. This is important because categorizing data

[1] The authors also mention this limitation on p.10, footnote #6.
[2] RR=response rate.
[3] MSPAP familiarity, support for MSPAP, current math instruction, and professional development

into discrete groups limits the variability that can be explained in any analyses. I also examine program effects over two years rather than one year, and include prior achievement as a predictor variable. I also use both student-level and school-level data, which may improve the validity of study findings.

Washington Assessment of Student Learning (WASL)

In 1993, Washington state's Education Reform Act mandated the creation of academic standards called the Essential Academic Learning Requirements (EALRs) (Stecher, 2010). The EALRs defined learning targets in a wide-range of subject areas (reading, writing, communication, mathematics, science, civics and history, geography, art, and health and fitness). The state assessment system, called the Washington Assessment of Student Learning (WASL), was developed to assess student mastery and proficiency relative to those content standards for reading, writing, math and science. WASL was implemented in 4th grade beginning in 1996 and included a combination of multiple choice, short-answer, essay, and problem-solving performance tasks (Stecher, 2010). Other grades were added so that by 2001, WASL was administered in reading and mathematics (grades 3-8, and 10), writing (grades 4, 7, and 10), and science (grades 5, 8, and 10). Individual, student-level scores were reported for school-level accountability purposes. WASL was replaced in 2009-2010 with the Measurements of Student Progress in grades 3 to 8 and the High School Proficiency Exam in grades 10 to 12 (Stecher, 2010).

Review of Stecher et al. (2000) Washington State Study

Stecher and his colleagues (2000) investigated the implementation and effect of Washington state reform on school and classroom practices in writing, reading, listening, and mathematics. Specifically, Stecher and colleagues investigated whether school practices significantly related to student achievement, controlling for school-level differences in percent free and reduced price

lunch, percent race/ethnicity, and school size. They did not include school-level prior achievement as a control.

To do so, the researchers administered a survey similar to the one used in Kentucky to a representative sample of about 150 elementary and middle school principals and about 400 fourth and seventh grade writing and math teachers from across Washington state during the spring of 1999. The schools were sampled using a stratified random sampling approach based on type of community in which the school was located (urban, urban fringe/large town, and small town/rural). Only middle schools that had voluntarily adopted the WASL were included in the sampling frame because WASLs were not required for 7th grade until spring 2001. Seventy-seven percent of principals surveyed responded (N=108 elementary and middle school principals) and sixty-nine percent of teachers responded (N=277 fourth and seventh grade teachers).

Stecher et al. (2000) analyzed the data using OLS multiple regression analysis to estimate 1) the effect of school demographics (percent free/reduced price lunch, percent race/ethnicity, school size) on school-level WASL scores (N=1401/subject area); and 2) the effect of school practices (as reported on the principal and teacher surveys) on school-level WASL scores in each subject area (reading, writing, listening, and mathematics), controlling for school-level variables (N=83 teachers/subject area). The researchers pooled 4th and 7th grade together in each subject area because of the low sample size.

Overall, there were a couple key findings. First, findings suggest that school-level demographics such as percent American Indian (B=-0.021, p<.001), percent free and reduced price lunch (B=-0.016, p<.001), percent Black (B=-0.015, p<.001), and percent Hispanic (B=-0.013, p<.001) have a negative effect on school achievement in all subject areas, typically in that order. On the other hand, percent Asian has a positive effect (B=0.017, p<.001) in all WASL subject area scores except in listening where it is non-significant. The parameter estimates just demarcated were

for reading and are typical of the other subject areas. The effects of school size and percent female were mixed, but also small. If effect sizes are calculated from the information provided in the appendices, the practical significance of these effects is arguably very little (for example, d=-0.0023 for the effect of percent American Indian on WASL reading scores). That said, in all subject areas, percent of free and reduced price lunch was the only significant school-level demographic predictor of WASL scores in models that included all the principal and teacher survey measures (B=-0.025, p<.001 in reading). These findings are not surprising given typical effects of socio-economic status on student achievement outcomes.

A second key finding is that few measured variables from the principal or teacher surveys are significant predictors of WASL scores, controlling for school-level demographic factors. For example, Stecher et al. (2000) found that reading (B=.169, p<.001) and mathematics (B=.138, p<.001) WASL scores were higher in schools where there was greater alignment between curriculum and the state standards as reported by teachers. This result, however, did not hold for the other two content areas (writing and listening). Mathematics scores were also higher in schools where teachers reported that they understood the state standards and assessment well (B=.279, p<.05).

That said, relationships between principal and teacher school practices and student achievement from the regression analyses were generally weak and unusual. For example, most variables had no significant relationship with WASL scores and the patterns of significant findings is sometimes in conflict. For example, they found a negative effect of curriculum alignment for listening (B=-0.18, p<.001), but positive effect for math (B=0.138, p<.001) and reading (B=0.169, p<.001). The findings are also sometimes counter-intuitive. For example, there was a negative effect of WASL-focused professional development on writing (B=-0.006, p=.05) and reading (B=-0.008, p<.05) scores, which seems unusual. Researchers state: "Such unusual results are not uncommon in regression analyses that include many variables that are correlated as these were" (p. 65). The

researchers did not provide a correlation matrix so it is impossible to estimate the extent of multi-collinearity between regression variables. It should also be noted that each regression model had a small sample size (N=83 teachers/subject area), which may limit the ability of the researchers to detect effects if one does exist in the population.

Review of Stecher and Chun's (2001) Washington State Study

Continuing the same line of investigation from the Washington state study the year prior, Stecher and Chun (2001) used OLS multiple regression analysis to investigate the relationship between 1998-99 and 1999-2000 school-level WASL scores with school practices, as well as principal and teacher perceptions as reported on surveys. The only difference in research design between this study and the one the year before (Stecher et al., 2000) is that stepwise regression was used to enter variables and separate models were specified at each grade level (4th and 7th) for each subject area. If variables are correlated this can be a problem for model building because the process may eliminate potentially important predictors. In the year prior, the multiple regression analysis was by subject area, but pooled the two grades together. Also, Stecher and colleagues provide reliability evidence on all study measures in the study's appendix.

Similar to findings in the prior study (Stecher et al., 2000), findings in general were inconclusive. The only variables that were significant predictors of WASL scores were aggregate student demographic factors such as percent free and reduce lunch and percent race/ethnicity. This may be an artifact of the stepwise regression procedure. Again, school mean prior achievement was not included as a control variable. There were a few cases where specific school or classroom practices were associated with higher WASL scores, but results were difficult to interpret just like in the prior study because some findings seemed counter-intuitive. For example, this study found a negative effect on 7^{th} grade writing scores when teachers reported taking more actions to support the reform ($B=-0.324$, $p<.01$). This finding is counter-intuitive because we would expect that the

more actions teachers take to support the reform would lead to positive effects on student achievement. These unusual findings led the researchers to conclude that they "did not find strong evidence that average practices measured by our surveys were directly related to school success on the WASL" (p. 24). In other words, the survey instrument itself was not sensitive enough to measure the school practices it was designed to measure.

Limitations and Implications of Washington State Studies

The conclusion of Stecher and Chun (2001) that their survey measures were not actually measuring what they were hoping to measure and only indirectly related to school success on WASL is a significant limitation of both studies just reviewed since they used the same survey items. The brief statement about potential multi-collinearity among the variables used in the regression analysis also raises questions about the validity of study findings. Furthermore, because the Kentucky research design and the Washington state research design were almost identical, these limitations suggests that the findings of the similar studies in those two states by the same researchers (Stecher et al., 1998, 2000; Stecher & Chun, 2001) may have found few to no relationships between reform-oriented practices (like the use of performance assessments) and student achievement because of research design limitations—not as a result of there being no relationships.

Given the limitations of survey methodology, a different research design may be warranted. For example, it may first be important and useful to establish the effect of the performance assessment program on student achievement outcomes and then do follow-up studies to investigate the contextual differences in program implementation such as changes in classroom practices that may explain those differences. This "backdoor approach" is agnostic to the differences in fidelity-of-implementation and changes in classroom-level practices so as not to assume that differences along those lines can be easily measured and used to predict differences in student achievement. This dissertation study, therefore, extends and adapts to difficulties experienced in prior research by

focusing on estimating the effects of an innovative assessment and accountability system on student achievement outcomes.

This study also extends and improves upon prior research by testing for the effects and cross-level effects of student- and school-level background and demographic characteristics on student achievement outcomes. Based on findings from these earlier studies, it is likely that student-level background characteristics such as free and reduced lunch (a proxy for socioeconomic status), as well as American Indian, Black, and Hispanic subgroup membership are likely to have negative effects on student achievement while Asian subgroup membership is likely to have positive effects. Also, this dissertation study extends this prior literature since it models the effect of student- and school-level prior achievement on current achievement.

There are a few other limitations of the two Washington state studies that I'd like to highlight in order to develop how my dissertation study aims to improve upon their research design. First, both studies aggregate school-level WASL scores to estimate program effects. Aggregate data limits variability and also ignores the nested structure of the data—that students are nested within schools. There was also a low sample size resulting from the teacher and principal survey responses. Only about 80 teachers per subject area responded in the Stecher et al. (2000) study, which is why the researchers pooled across the two grade levels.

Another limitation of the Washington state studies is that the WASL implementation timeline was gradual and done over a 10-year period of time. Only 4th grade WASL testing was required during the two study's timeframe (1998-2000). As mentioned earlier, 7th grade WASL testing was not required of districts until spring 2001. The middle schools used in each study were voluntarily administering the WASL and it may be that those schools are in some way different than other non-early adopting schools in the state and those differences are related to student achievement outcomes as measured on WASL. This selection bias may affect study findings. My

study addresses this limitation because it uses an outcome measure in common and required of all schools in the state.

Maryland School Performance Assessment System (MSPAP)

Maryland's School Performance Assessment System (MSPAP) was sparked by a 1989 report that called for improved student achievement and academic standards (Stecher, 2010). According to Parke, Lane and Stone (2006), the goal of standards-based reform was to influence curriculum, instruction, and assessment, "and the use of performance tasks on assessments were considered to be an integral part of the reform" (p. 240). First administered in 1991, MSPAP was developed to measure student proficiency on the Maryland Learning Objectives (MLOs). MSPAP assessed six subjects (reading, writing, language, math, science, and social studies) in grades 3, 5, and 8 (Lane, Parke, & Stone, 2002). The MSPAP contained *only* open response, performance tasks that ranged from simple to more complex; some tasks were multi-disciplinary and entailed group work (Koretz et al., 1996). The goal of the MSPAP was to drive the use of performance-based instruction at the classroom level (Parke, Lane, Stone, 2006).

MSPAP was designed as a school accountability system. Matrix sampling was used, which means that the items were sampled so that every student only took a portion of the exam in each subject (Stecher, 2010). Results were reported for schools and districts and there were rewards and sanctions for schools based on the results (Parke et al., 2006). Teachers received professional development from the state to support the implementation of MSPAP. MSPAP lost political support due to concerns with scoring and wide fluctuations in school-level scores from year to year (Stecher, 2010). Also, because of the requirements of NCLB where individual scores needed to be reported, MSPAP was replaced in 2002. It was replaced by a test in reading, math, and science that was mainly multiple-choice with some constructed response (Stecher, 2010).

The following three articles report on the same research study. These studies examined the effects of the MSPAP and Maryland Learning Outcomes on curriculum, classroom instruction and assessment practices, professional development activities, and student learning. The same sampling techniques, research questions, and research designs are used in each article, although the final article also analyzes a collection of classroom instruction and assessment materials. Each article reports results for a different grouping of subject areas. The first study reports on math (Lane et al., 2002). The second study reports on five subject areas: math, reading, writing, science, and social studies (Stone & Lane, 2003). The third and final study on reading and writing (Parke et al., 2006). In most cases, the research includes data from 1993-1998—a five-year time span—which are not the first five-years of implementation. MSPAP was first implemented in 1991 so this research does not capture the first two years of program effects.

I chose to review these articles last even though the research time frame overlaps other research conducted in Kentucky, for example, because these articles were published later and they are more methodologically advanced. The researchers state that some of the items from the survey instrument they developed pertaining to support and beliefs about MSPAP were based on a study of the perceived effects of the MSPAP conducted by the same researchers that completed the Kentucky and Washington state studies (Koretz, Mitchell, et al., 1996). This suggests that some of the limitations of the surveys in those studies (i.e., lack of measurement sensitivity) may also pertain to these Maryland studies, as there was some cross-pollination of survey items. The overarching study in Maryland was set in the context of a consequential validity argument whereby the high-stakes nature of the assessment program means that the uses and interpretations of the assessments need to be addressed, including the "(a) negative and positive consequences and (b) intended and plausible unintended consequences" (p. 2).

Review of Lane et al. (2002) Study

Lane, Parke, and Stone's (2002) study included 90 schools (59 elementary schools; 82% RR[2] and 31 middle schools; 86% RR) in their sample based on the same random stratified sampling method (percent free and reduced lunch and MSPAP performance gains) used in all three Maryland studies. They administered questionnaires to grade 2-8 principals, teachers, and students during the 1996-1997 school year on a range of dimensions related to math curriculum, classroom instruction and assessment practices, and professional development. Different from the surveys administered in Kentucky and Washington state, the researchers used confirmatory factor analysis to validate the factor structure of the questionnaires. This study also used advanced statistical techniques (latent variable growth modeling) to model the growth in average school-level math performance on MSPAP over 5 years (1993-1998) and how that growth related to responses on the questionnaires. They also examined how the effects varied by grade levels (MSPAP-on grades: 3, 5, and 8; and MSPAP-off grades: 2, 4, 7), as well as school characteristics such as percentage of students who qualify for free and reduced lunch.

Lane, Parke, and Stone's (2002) study had two major findings. First, they found that the percentage of free and reduced price lunch students in the school correlated significantly with both the initial 1993 school-level MSPAP math performance and 1997 school-level MSPAP math performance, but not with the slope (or rate of change). This means that the percent of students in a school who qualify for free and reduced lunch (a proxy for socioeconomic status) is not related to MSPAP performance gains just initial and final performance. This finding is similar to the Washington state studies that also find a negative effect of percent free and reduced price on school-level WASL scores.

[2] RR=response rate.

And, second, the study found that the only factor out of five that explained variability in school-level rates of change was the "MSPAP Impact on Instruction" dimension on the teacher questionnaire.[3] Higher levels of teacher reports of MSPAP "having a direct impact on instruction" were associated with higher rates of change in MSPAP school performance over five years ($B=-1.1$, $p<.05$)[4](p. 310). This finding is similar to the results of the Stecher et al. (2000) study in Washington state. In that study, schools where teachers reported greater alignment between the curriculum and state standards also had higher math scores.

Lane et al. (2002) go on to argue that the Linn, Baker, and Betebenner's (2002) analysis of trends in percentages of 8th grade students meeting the performance standard on MSPAP in math from 1994 through 2001 and the trends in the percentages of 8th grade students at basic or higher performance levels on NAEP math assessments in Maryland from 1990 to 2000 were "very similar" (p. 313). Lane, Parke, and Stone (2002) argue this provides evidence that the MSPAP gains are replicated on other assessments (e.g., NAEP) so that the teacher-reported changes to instruction were "not superficial changes to increase performance on MSPAP but were more substantive changes that enhanced students' understandings in mathematics" (p. 313). The researchers likely argue against "superficial changes" because it addresses what some consider to be potentially invalid score inflation and score gains due to "teaching to the test" rather than actual improvement in student learning (e.g., Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz, 2005).

[3] MSPAP familiarity, support for MSPAP, current math instruction, and professional development support were the other four factors that did not significantly predict or explain variability in school-level rates of change on MSPAP.

[4] Because the rate of change is scaled from 1997 to 1993 (in reverse), the negative parameter estimate means that higher levels of MSPAP impact on instruction are associated with greater rates of decrease from 1997 to 1993 (or higher levels of rate of change) in MSPAP school performance.

Review of Stone & Lane (2003) Study

This article by Stone and Lane (2003) reports on five subject areas: math, reading, writing, social studies, and science. Using the same 90 elementary and middle schools (grades 2-8) detailed in Lane et al. (2002) selected in the 1996-1997 school year to report on math and now reading and writing, an additional 161 schools were selected in the 1998-1999 school year to report on science and social studies in this article. Due to the small sample sizes, the data for elementary and middle schools were pooled.

There are two main findings from this study. First, Stone and Lane summarized the mean MSPAP performance across schools in the sample over the 5-year time period (1993-1998). The general trend was a significant increase in mean performance over the 5-year time period. For example, except with writing, there were larger gains in the early years, followed by a leveling in the middle year, with an increase in mean performance for schools in the last two years. In writing, there was a slight dip in the early years followed by a steady increase over the last three years. This finding suggests that it may be possible to detect program effects at least by Year 3 or 4 of a performance assessment program implementation. The study cannot provide evidence about the first two years of MSPAP since it did not include those years in the study.

Second, similar to the previous Maryland article that reported just on math, they found that the percent of students in a school who qualify for free and reduced lunch was related to initial performance levels in all subject areas, but not rates of change over time except reading where there was a small negative effect ($B=-0.03$, $p<.05$). For example, increases in the percent of students who qualify for free and reduced price lunch were associated with lower levels of MSPAP performance in 1997 or 1998 ($B=-0.6$ in reading, $p<.05$; $B= -0.8$ in math, $p<.05$).

However, similar to the previous studies in Kentucky and Washington state that also explored relationships between survey responses and student achievement, there were some unusual

and counter-intuitive findings. For example, increases in student motivation levels were associated with decreased MSPAP initial performance in social studies ($B=-5.7$, $p<.05$) and rate of change in science ($B=-2.1$, $p<.05$), but non-significant in all the other subject areas. Also, higher levels of teachers reported use of reform-oriented tasks was negatively associated with MSPAP performance gains over time for both writing and reading ($B=-2.4$, $p<.05$ and $B=-1.7$, $p<.05$, respectively). The researchers do not easily explain these unusual patterns and counter-intuitive findings. This again suggests that another research study approach to exploring performance assessment program effects on student achievement may be needed.

Review of Parke et al. (2006) Study

Using the same sample and 5-year time frame (1993-1998) from Lane, Parke, and Stone (2002) that reports on math and Stone and Lane (2003) that reports on five subject areas, the same researchers highlight their findings in reading and writing. The key difference in this study by Parke, Lane, and Stone (2006) is that the researchers also selected a random sample from the 90 elementary and middle schools in the two earlier studies and asked those schools to participate in the collection of classroom instruction and assessment materials. This methodology represents an expansion of previous methodologies in this line of research towards more contextual classroom-level information and mixed-methods approaches. This methodological expansion addresses some of the inherent limitations with self-reported survey responses and the extent to which teacher perceptions are lived out in actual classroom practice.

Forty-four out of the fifty-one schools randomly sampled agreed to participate, representing 15 of the 24 school counties in Maryland. In the 44 schools, 280 reading and writing teachers sent in a total of 3,221 classroom activities and 1,296 classroom assessments. Two coding schemes were developed based on the Maryland Learning Outcomes (MLOs) and MSPAP performance task format and content to examine the large number of classroom artifacts (about 4,500) that were

collected. Researchers coded the artifacts on dimensions related to level of alignment with MSPAP and MLOs, response type required of students, level of integration with other subject areas, and amount of group work, for example.

Because the other parts of this study are exactly the same as what was reported in the Stone and Lane (2003) article that I just reviewed, I do not review study findings related to MSPAP performance gains in reading and writing and how initial performance gains and rates of change were related to survey responses. Instead, I focus on findings related to the collection of classroom artifacts and how they add to what is known and foreshadow what is still left to understand regarding the effects of a performance assessment program on student achievement.

One key finding from this artifact investigation was that teachers over-reported the use of classroom practices aligned to the MSPAP and MLOs. For example, the average teacher response to questions about their current reading and writing instruction was a "3" on a 4-point Likert-type scale ranging from "no alignment"=1 to "a great deal of alignment"=4. However, the artifact analysis revealed most activities were low on MSPAP and MLO alignment. Interestingly, the student survey responses were a more accurate reflection of the degree of alignment. These findings suggest that asking teachers to report on their use of reform-oriented practices may not be an accurate reflection on actual classroom practices and therefore less helpful in terms of explaining or predicting differences in student achievement. Parke, Lane, and Stone (2006) say as much themselves as they mention discrepancies in other research on teachers' self-reported practices and "potential discrepancies between the intended curriculum, the implemented curriculum, and the attained curriculum" (p. 263). These findings also suggest that the benefits of artifact analysis pertain especially to exploring the level of alignment between the reform objectives and actual classroom practices; however, one of the limitations of artifact analysis is that even if there is a high level of

alignment, it is still unclear what effect that alignment has on student learning over time unless it is paired with an examination of student achievement trends like it is in this study.

Limitations and Implications of Maryland Studies

There are a few limitations of the Maryland study reported in the three articles reviewed above. First, the researchers were limited to a school-level analysis because the MSPAP uses matrix sampling. This means the researchers were not able to estimate differences in student achievement within- or between-schools due to student-level factors or cross-level effects. Also, the researchers pooled elementary and middle schools together in their analyses due to small sample sizes. Findings might have been different had they examined the grade spans separately. The small sample size also inhibited the inclusion of many of the teacher, principal, and student survey dimensions in the latent growth model analysis.

Overall, however, the Maryland study has a strong research design that flows from the research purpose and research questions. The researchers expanded upon the questions asked in concurrent studies in Kentucky to explore the extent to which differences in elementary and middle school performance *over time* are associated with teachers' reported changes in classroom practices. For example, the researchers examined effects over five years, which is a significant expansion from the one-year investigations in all the other studies. The researchers also took this line of research in a new direction by collecting and analyzing classroom artifacts. They used these classroom artifacts to explore the extent to which teacher's self-reported use of reform-oriented practices was reflected in *actual* classroom activities and assessments.

The 5-year pattern of MSPAP school performance gains provides some evidence that it may be possible to detect performance assessment program effects as early as Year 3 or 4 after implementation. No evidence is available on Year 1 or 2. The general pattern was a sharp gain for two years, followed by a leveling effect for one year, with another increase in mean performance for

schools in the last two years. Writing was the only exception because there was a dip in the first year of the study (which was actually Year 3) followed by a steady increase across the other four years of the study time frame.

However, there are some key differences between the Maryland performance assessment reform and the reform under investigation in this dissertation. For example, the Maryland reform aligned the goals of the program with the state test format (i.e., only performance tasks) so that program effects may be more apparent on the MSPAP in a shorter amount of time. In this dissertation, the state test format is not completely aligned with the goals of the program, although the achievement test does include performance tasks. Thus, program effects in this dissertation may be more indirect in state-level testing years and take time to accumulate in order to see effects on student achievement. Furthermore, the Maryland study did not examine program effects in the first two years of implementation so it is unclear how the performance assessment program may have influenced student learning during early implementation.

Perhaps one helpful way of using the Maryland results (given the caveats just mentioned and the fact that elementary and middle school results were pooled) is to use the standardized mean differences in school performance over the first two years of the study (1993-1995) by subject area as an upper and lower bound on expected school performance growth after two years. For example, MSPAP school performance in math increased by 0.32 SD-units between 1993-1995 and reading by 0.50 SD-units, which may be considered upper bounds. Because writing exhibited nonlinear growth and dipped between 1993-1994 before increasing steadily, MSPAP school performance in writing increasing 0.07 SD-units over two years can serve as a lower bound.

Another implication of the Maryland study is that there are key limitations with teacher self-reported survey data in this context. The researchers encountered similar problems that faced other researchers in Kentucky and Washington state—how to make sense of the unusual patterns and

counter-intuitive findings that result when trying to explain differences in student achievement with teacher self-reported data. For example, in the Maryland study there was no teacher, principal, or student survey variable that explained initial school performance levels or rates of change over the five years in every (or even most) subject areas. Furthermore, findings that there were potential discrepancies between what teachers self-report and actual classroom instruction and assessment practices (i.e., teachers over-report their use of reform-oriented practices) is sobering for anyone interested in using survey data related to classroom practices to explain changes in student achievement.

Synthesis Across Performance Assessment Program Studies

This synthesis traces the development of particular issues/themes in the research across performance assessment program studies. This synthesis focuses on what is known and what is left to understand about the effects of school-, district-, or state-level performance assessment programs on K-12 student achievement outcomes in math and ELA. Since there are seven studies in this research area and three of those report on the same study, there is a lot left to understand. I interweave this section with the main implications of these studies for this dissertation study as they arise.

There are at least three issues/themes that can be traced through the research on performance assessment programs. First, *most of the research in this area did not focus exclusively on examining the effects of a performance assessment program on student achievement outcomes*—the Shepard and colleagues' (1995) study being the exception. Instead, the research focused on examining how teachers' self-reported changes in classroom practices from implementing a performance assessment program related to differences in student achievement. These studies used surveys to examine teacher and sometimes also principal and/or student perceptions of changes in classroom practices. Researchers were particularly interested in understanding how the implementation of a new

assessment and accountability system that utilized performance-based assessments either solely (as in Maryland) or in conjunction with multiple choice and constructed response (as in Washington state and Kentucky) may have influenced classroom practices and therefore student achievement. The theory of action for those performance assessment programs was the same as the one under examination in this dissertation; namely, reforming assessment can drive better teaching practices that then improves student learning. Overall, researchers sought to answer the question: Which teaching practices were more strongly related to improvements in student learning than others?

Given that research focus, findings across these studies were generally inconclusive, although there is some evidence to suggest that schools where teachers reported greater alignment between the curriculum and state standards also had higher math scores (Lane et al., 2002; Stecher et al., 2000; Stecher & Chun, 2001). However, there is clear evidence across studies that there is a negative effect of socioeconomic status (percent of students in the school who qualify for free and reduced lunch) with average school performance in all subject areas. School-level socioeconomic status does not appear to effect rates of change in school performance over time, except potentially in reading where there was a very small negative effect in the Maryland study (Stone & Lane, 2003). This implies student- and school-level free and reduced price lunch is an important control variable in any future study in this research area.

The difficulty faced at least to some extent by all the researchers who used surveys, however, was making sense of unusual or nonsensical results. This led the Kentucky and Washington state research team to question the sensitivity of their survey instrument—some items of which were used by the Maryland research team. Another threat to the internal validity of study findings highlighted by the Maryland research team is the potential discrepancies between teacher self-reported practices and actual classroom activities and assessment practices. Teachers may over-report their use of reform-oriented practices for many reasons, including memory effects and social desirability bias.

This suggests that the widely favored survey research design might adversely affect study outcomes in this context and that another approach to exploring performance assessment program effects on student achievement may be needed.

Second, *there is some evidence to suggest that performance assessment programs may have a small positive effect on student achievement in both math and ELA over time*. For example, in the two studies that examined the effects on student achievement outcomes directly, one found a small positive effect on one outcome measure in math after one year (d=0.13), but no effect on either outcome measure in reading after one year (Shepard et al., 1995). The other study found a significant increase in average school performance over five years in five subject areas, including: math, reading and writing (Stone & Lane, 2003). A key difference between these studies is the number of years included in the analyses. One year of data may be too little to see evidence of performance assessment program effects, but that it is possible to detect effects after two years. It is unclear from the research literature if program effects are evident after two years of implementation.

Third, *none of the prior research modeled dosage effects, allowed non-linear treatment effects, or examined differential effects for certain subgroups of students*. My dissertation research aims to fill these gaps in the research literature on performance assessment programs, as well as the gaps that are discussed next in the competency-based education literature.

Background on K-12 Competency-Based Education in the United States

Competency-based education—also known as proficiency-based, mastery-based, and performance-based education—has no clear cut definition, but typically has at least these four defining features: (1) students advance upon mastery, (2) students receive support and progress monitoring based on their individual learning needs, (3) the content and assessment of student learning is flexible and personalized, and (4) school policies and structures support anytime/anywhere learning (CompetencyWorks, 2014; Le et al., 2014). Students advance upon

mastery means that students move on when ready or progress in the curriculum through demonstration of mastery not just how many hours or days they spent in a classroom. In other words, students must demonstrate that they have learned what was expected before moving on to new material. Demonstrating proficiency upon readiness rather than adherence to the Carnegie unit is fundamental to the concept of competency-based education. Students do not progress in the curriculum based on the amount of time they spend in school, but based on mastery of the material. This requires flexible pacing and flexible structures such as placing students in classes based on their level of understanding rather than their age/grade level. This also requires sophisticated support structures and progress monitoring so that students are provided personalized and customized support based on their learning needs. Personalization of content refers to student choice in the content of their learning and how learning is delivered and assessed. Flexible assessment of student learning refers to the timing of assessments and the types of assessments used to determine student proficiency. This is intended to allow students more choice and voice in their learning goals and how they provide evidence of proficiency with the goal of more student engagement.

The rationale behind the competency-based education movement has its roots in the progressive education movement in the early 1900s (i.e., John Dewey)(Le et al., 2014). The goal of competency-based education is to reduce inequities in student achievement outcomes and achievement gaps. The underlying premise is that the problem with the traditional system is that students are passed on from one grade to another, even if they have not mastered the content (i.e., social promotion). This is how we find high school students who have progressed through elementary and middle school, but still don't know how to read on grade level.

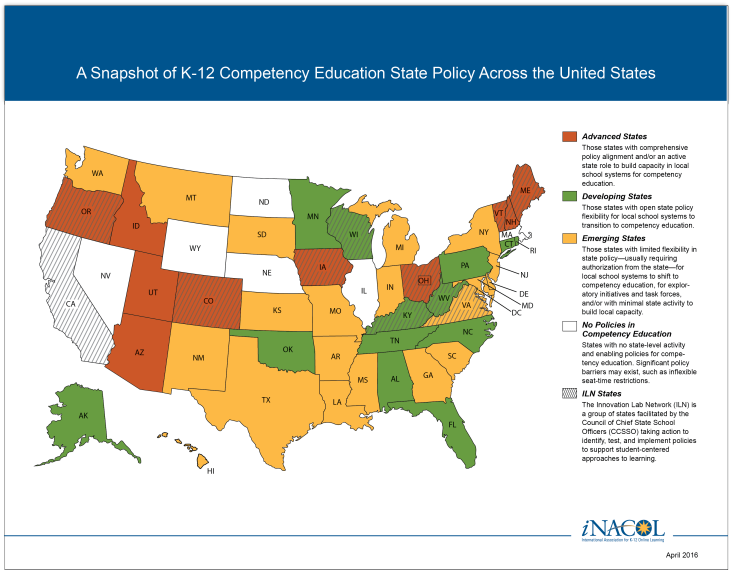
The competency-based education movement builds upon Benjamin Bloom's concept of "mastery learning" and the belief that all students can master the academic standards as long as instruction is tailored to their learning needs and they do not proceed to other concepts until they've

mastered pre-cursor concepts (Bloom, 1968). For this reason, performance-based assessments and other assessments that utilize rubrics and qualitative descriptions of performance are often incorporated into competency-based education so that students know what they need to do in order to master the content both before and after the assessment.

In the last 10 years, competency-based education has once again caught the eye of education reformers looking for a way to reshape educational systems to ensure all students "reach proficiency in the skills they need for college and careers" (CompetencyWorks, 2011). Applying Downs' (1972) issue-attention cycle, a recurring pattern of connected ideas and policy solutions emerges: John Dewey's child-centered movement in the early 1900s, Benjamin Bloom's mastery learning movement in the 1970s and 1980s, and the more recent competency-based education movement. Although both the child-centered and mastery learning movements were not fully implemented at scale, some teachers developed a hybrid approach between the traditional model of education and the "newer" pedagogical practices (Kliebard, 2002). Furthermore, many educator preparation programs today teach future educators about John Dewey and Benjamin Bloom so their ideas and practices are still present in American education today albeit in a diluted form. The newer competency-based education movement aims to do what neither Dewey or Bloom were able to do—reconfigure patterns of teaching and learning and school structures so that time is viewed more flexibly.

According to Chris Sturgis, a national leader in competency-based education, there are ten "advanced" competency-based states because the states have "comprehensive policy alignment and/or active state role to build capacity in local school systems for competency education" (see Figure 2.1)(Sturgis, 2016). These states include: New Hampshire, Maine, Vermont, Ohio, Iowa, Colorado, Utah, Arizona, Idaho, and Oregon.

Figure 2.1 Overview of K-12 competency-based education policies in the United States as of April 2016



Another thirteen states are deemed "developing" in terms of competency-based education models because they have "open state policy flexibility for local school systems to transition to competency education" (Sturgis, 2016). However, according to Chris Sturgis, there are many more schools and school districts all over the United States that have transitioned or are transitioning to a competency education learning environment to close achievement gaps and help all students attain college or career readiness before graduation from high school (Sturgis, 2016).

Review of the Research Literature on the Effects of Competency-Based Education on Student Achievement

This review of the research literature on the effects of competency-based education on student achievement is organized into two main sections. The first section reviews the research literature from the 1970s and 1980s on mastery learning. The second section reviews the research literature on the more recent resurgence of competency-based education (2005 to present).

Most of the research on competency-based education is from the 1970s and 1980s. During that time, competency-based education was referred to as mastery learning. I reviewed only meta-analytic and meta-synthetic studies related to the effects of mastery learning on K-12 student

achievement outcomes. I did not include research reviews on the effects of postsecondary mastery learning programs (e.g., Kulik et al., 1979). I focus on K-12 research reviews and do not drill down to review the hundreds of individual mastery learning studies during that time period because the competency-based education movement taking place in the last ten years in the United States is similar to, but not exactly the same as the mastery learning movement from the 1970s and 1980s. For example, mastery learning obviously emphasized mastery (or progress upon demonstration of proficiency), but depending upon the type of mastery learning program, the pacing of instruction varied. Also, there wasn't an emphasis in mastery learning programs on personalized learning or instruction and it was prior to the standards-based movement that started in the 1980s. As a result, reviewing the effects of mastery learning on K-12 student achievement outcomes may be only instructive for this study to a certain point. Reviewing the meta-analytic and meta-synthetic findings provides a general background of what is known and what is left to understand, which the more recent studies then build upon.

It is important to note that in the second section, any study that examined the effects of an educational intervention with all four defining features of competency-based education (students advance upon mastery, support and progress monitoring, personalized content and assessment, and flexible school policies about where learning takes place) on K-12 student achievement outcomes was included in this review, even if the title of the study or research report did not say "competency education." This is because competency-based education has a broad definition and, according to CompetencyWorks (2011), many terms can fall under the umbrella of competency education such as "proficiency-based," "mastery-based," or "performance-based." For example, a study on personalized learning was included (Pane et al., 2015) because the four elements were present, but a study on deeper learning was excluded (Zeiser, Taylor, Rickles, Garet, & Segeritz, 2014) because the four elements were not present. Three separate studies were located that more recently examined K-

12 student achievement outcomes associated with competency-based education reforms (Haystead, 2010; Pane et al., 2015; Steele et al., 2014).

This review is organized chronologically beginning with the K-12 meta-analytic and meta-synthetic research on mastery learning reforms from the 1970s and 1980s and then moving to the competency-based reforms from the last decade. First, however, a quick overview and explanation of mastery learning follows in order contextualize the meta-analytic and meta-synthetic research findings below.

Mastery Learning Research Reviews

Mastery learning is a theory about teaching and learning based on the ideas of Benjamin Bloom (1968) that prescribes certain instructional strategies (Block & Anderson, 1975). It is important to note that there were a wide variety of programs that fell under the umbrella of mastery learning in the 1970s and 1980s. For example, mastery learning can either be individualized or group-based. In individualized mastery learning approaches, students can move at their own pace; whereas, in group-based mastery learning approaches, students who demonstrate mastery over the material pursue enrichment activities until most of the class is ready to move on and start a new learning activity together (Cotton & Savard, 1982).

Slavin (1987) describes three primary forms of mastery learning: Keller's Personalized System of Instruction (individually-paced; postsecondary level); continuous progress (individually-paced; K-12); and Learning for Mastery (group-based; K-12). As stated earlier, the primary premise of all mastery learning approaches is the basic belief that all students can learn "when provided with the conditions that are appropriate for their learning" (Guskey & Gates, 1986, p. 73).

According to Guskey (1986), there are two crucial elements to any mastery learning program: (1) feedback and correctives and (2) congruence among instructional components. Feedback and correctives means that mastery learning programs included formative assessment practices that

helped students identify key learning targets, how well they have learned those targets, and where additional time and/or re-teaching needed to occur. The second component was the consistency and alignment between the educational objectives, instructional strategies, and formative assessment practices. For example, if students are expected to learn critical thinking skills, mastery learning specifies that the instructional strategies used to teach students those skills allow students to employ critical thinking and then students are given specific feedback on their critical thinking skills alongside directions for how to correct and improve. Guskey (1994) argues that Bloom did not prescribe any particular curriculum, instructional method, or assessment form—in essence his theory was neutral on those topics—just that there was consistency and alignment between the three components.

Review of Block and Burns's (1976) Synthesis

Block and Burns (1976) synthesize the mastery learning research using two methodological inclusion/exclusion criteria. First, studies had to have a substantial degree of external validity, which they define as "performed in an actual school setting and employ school-like learning tasks...that were meaningful, complex, and relatively long" (p. 13). Second, the research study had to have a substantial degree of internal validity. That is, only studies where experimental groups were roughly equivalent at baseline because of the use of a randomized or quasi-experimental design were included. Block and Burns then sorted the research into four categories or types. Type 1 studies focused on the question, "Does mastery learning work?" and compared the learning of mastery-taught students to the learning non-mastery-taught students using performance on the same end-of-course examinations. Type 2 studies focused on the question, "If mastery learning works, then what might follow?" and tended to examine the affective consequences of mastery learning. Type 3 studies examined why mastery learning approaches work. Type 4 studies examined how mastery

learning approaches work. This review focuses on the Type 1 studies since those are the only studies that examine the effects of mastery learning on K-12 student achievement outcomes.

Block and Burns review thirty-eight studies in the Type 1 group, most of which have relatively short intervention periods (2 weeks to 16 months). These studies include both individualized and group-based mastery learning approaches, many different subject areas, and grade levels (elementary through postsecondary). They sort the studies by type of mastery learning approach—individualized or group-based—as all individualized approaches they found focused on postsecondary students. I include findings only on the group-based mastery learning approaches since I'm interested in effects on K-12 student achievement. The researchers examine findings related to how well mastery students perform in comparison to non-mastery students, as well as how much variability the students exhibit in their learning. Block and Burns argue that if mastery learning approaches help students learn better than not only should students have higher achievement, but they should exhibit less variability in their achievement.

The researchers found that group-based mastery learning students tended to score around 0.83 standard deviations higher than non-mastery learning students on locally constructed tests. They did not compare performance on standardized achievement tests because researchers tended to create their own dependent measures. They also found mastery students exhibited less variability in their performance than non-mastery students about 75% of the time. Overall, Block and Burns argue that group-based mastery learning approaches had positive effects on K-12 student achievement.

Limitations and Implications of Block and Burn's (1976) Meta-Synthesis

Block and Burns (1976) stipulate three limitations of all mastery learning studies to this point in time: (1) the use of locally constructed dependent measures of student performance without adequate description and detail about how those measures are constructed or evidence surrounding

their validity and reliability; (2) the mastery learning treatment is not adequately described; and (3) the non-mastery treatment is not adequately described. To this list, it also appears from Block and Burns's brief descriptions of the studies that many studies do not include an equivalent control group, many studies only use a posttest design, and the studies only report on short-term outcomes.

Given these limitations, it is difficult to ascertain the impact of selection bias on study findings, longer-term outcomes of mastery learning on student achievement, as well as the accuracy of the dependent measures in measuring what they purport to measure. Another limitation of the synthesis is that it provides no information on how mastery learning programs vary in their effectiveness based on grade level or subject area. These limitations underscore the importance of my dissertation study design, particularly around choosing the dependent measure, creating equivalent treatment and comparison groups at baseline, investigating effects by grade level and subject area over two years, and describing the treatment and non-treatment conditions in detail.

Overall, Block and Burns's meta-synthesis suggests that relatively short duration, group-based mastery learning interventions tend to have a small positive effect on K-12 student learning on locally constructed measures of student performance across multiple subject areas.

Review of Cotton and Savard's (1982) Meta-Synthesis

Cotton and Savard (1982) synthesize findings from thirty-three resources (24 primary research studies and 9 secondary research reviews). In total, they reviewed over 100 studies and evaluations of individualized and group-based mastery learning from elementary to post-secondary. Most of the studies they review pertain to elementary and secondary students (26 out of the 33 studies) and cover a wide range of subject areas—math, science, reading/language arts, social studies, etc. They report on effects of mastery learning related to student achievement, retention and attitudes. This review focuses only on student achievement.

The researchers rated each resource on a scale from 1-5 based on the quality of the study. They also categorized studies based on the hypothesis that "the use of mastery learning strategies with elementary and secondary students produces achievement results superior to those resulting from non-mastery instruction" (Cotton & Savard, 1982, p. 15). They categorized resources into those that tended to support the hypothesis, resources that tended to deny the hypothesis, resources that are inconclusive regarding the hypothesis, resources which were excluded because they were weak, and resources which were excluded because they were judged to be irrelevant to the hypothesis. Overall, they found that 23 resources tended to support the hypothesis, 7 resources tended to deny the hypothesis, and 3 resources were deemed irrelevant. The only additional information provided by the meta-synthesis is that the 23 resources that supported the superiority of mastery learning strategies on student achievement outcomes spanned many subject areas, both elementary and secondary grade levels, and different student aptitude levels. Cotton and Savard point out that "several of the researchers noted that low-aptitude students benefitted even more than other students from this instructional approach" (p. 7).

Limitations and Implications of Cotton and Savard's (1982) Meta-Synthesis

The overarching limitation of the Cotton and Savard (1982) meta-synthesis is that it provides a broad overview of the effectiveness of mastery learning without many details. For example, there is no explanation as to why the 7 resources they describe as "well-structured studies" (p. 8) failed to detect any effects of mastery learning on student achievement. It would have been helpful if the researchers had reviewed those 7 study designs in detail, especially in comparison to the 23 resources that did find effects to note any major differences in research design, population, sample size, outcome measure, and/or methodology that might explain differences in findings. For example, how did each study measure student achievement? Were the outcome measures researcher-created or standardized achievement tests?

In addition, it would also have been helpful if the researchers explained effects by grade span, subject area, and type of mastery learning approach (individualized vs. group-based) instead of just overall findings. That said, findings from Cotton and Savard (1982) provide more evidence that mastery learning has positive effects on K-12 student achievement outcomes in a range of subject areas than the alternative hypothesis.

Review of Guskey and Gates's (1986) Meta-Analysis

Guskey and Gates (1986) utilize meta-analytic techniques to synthesize the research on group-based K-12 mastery learning programs. Although they focus on a wide-range of student outcomes (such as student retention, student affect, and student achievement), The review focuses on student achievement outcomes. The researchers used three main inclusion/exclusion criteria to cull the literature. First, only studies between 1975-1985 were included for review because they thought Block and Burns's (1976) meta-synthesis provided a comprehensive review of the literature prior to 1975. Second, they only included studies that examined group-based mastery learning approaches. Third, the studies had to report data on measured outcomes for both treatment and control students and be free from serious methodological flaws. This limited the research to 38 studies and they narrowed this down to 27 studies as they chose to focus on only elementary and secondary classrooms. Two of those 27 studies did not focus on student achievement, but other student outcomes so there was a total of 25 studies reviewed related to K-12 student achievement outcomes. As in the Block and Burns's (1976) study, the most common measure of student achievement was student scores on teacher created unit- or end-of-course examinations.

There are two main findings from Guskey and Gates's (1986) meta-analysis of group-based mastery learning on student achievement outcomes. First, mastery learning students outperformed non-mastery learning students in every one of the 25 studies. However, the researchers chose not to calculate an average effect size across the 25 studies because there was so much variability from

study to study. For example, effect sizes ranged from 0.02 in one study to 1.7 in another study. To examine possible reasons for the significant variation in study effect sizes, Guskey and Gates examined results along two dimensions: grade level and subject area. This leads to the second major meta-analytic finding. They found that elementary and middle school studies exhibited larger mean effect sizes in comparison to high school studies (ES=0.89, 0.93, and 0.72, respectively). They also found that effects in math and science were weaker than effects in social studies and language arts (ES=0.78 science; 0.81 math; 0.91 social studies; 0.99 language arts). It is important to note that these effect sizes are on teacher created tests. To put these effect sizes into context, the strength of these effect sizes is similar to those found by studies investigating the effects of feedback and formative assessment on student achievement (Hattie, 2009).

Limitations and Implications of Guskey and Gates's (1986) Meta-Analysis

The limitations of Guskey and Gates's (1986) meta-analysis are similar to those of Block and Burns (1976). For example, the dependent measure used in most of these studies were teacher-created unit- or end-of-course examinations and not standardized achievement tests. These teacher-created assessments are not well described and the evidence of their validity and reliability is not provided. It is unclear how study findings would have been different in either synthesis if a standardized achievement test had been used as the dependent measure.

Furthermore, as Guskey and Gates explain, variability in the magnitude of effects may also result from how mastery learning is defined and implemented in each study. They state: "there is confusion and debate as to what is, and what is not, mastery learning" (p. 79). Since many of the studies do not include detailed descriptions of the mastery learning treatment or non-mastery control, it is difficult to disentangle treatment effects from other sources of variation in the treatment and control groups. This again has implications for this dissertation study design. Overall, Guskey and Gates's meta-analysis suggests that there are positive effects of group-based mastery

learning on K-12 student achievement outcomes across grade levels and subject areas, but there is considerable variation in effect sizes.

Review of Slavin's (1987a) Best-Evidence Synthesis

Robert Slavin may be best known for his best-evidence synthesis approach for synthesizing large literatures in the social sciences (1986). This approach combines features of meta-analytic and traditional narrative reviews. Slavin's inclusion and exclusion criteria differ significantly from prior syntheses of mastery learning research. For example, he includes only group-based mastery learning approaches in elementary and secondary schools that take place over periods of at least 4 weeks. Excluding studies with durations less than four weeks removed many studies that had been included in previous reviews and meta-analyses (e.g., Block & Burns, 1976; Guskey & Gates, 1986). Slavin argued this inclusion criteria was necessary because he was interested in examining effects on student achievement in practice, not just in theory. Mastery learning is intended to be an instructional strategy used over the course of the year, not just in a limited window of time. Also, Slavin only included studies if they provided evidence that the treatment and control groups were equivalent at baseline, or the degree of nonequivalence was reported so that effect sizes could be adjusted. Slavin's best-evidence synthesis included a total of 17 studies.

Slavin analyzed the research literature based on claims. First, he examined studies that provided evidence about the "strong claim;" that is, mastery learning is more effective than traditional instruction even when time is held constant and both content coverage and mastery are measured. He reviewed seven studies that adhered to the inclusion/exclusion criteria, held time constant, and used a standardized measure of student achievement (not a teacher- or researcher-made test). He found that the effects of group-based mastery learning on standardized achievement tests were "extremely small, at best" (p. 187). The median effect size across the 7 studies was 0.04,

which Slavin calls "essentially zero" (p. 187). This is a significantly different finding from prior research in this area.

To explain potential reasons for why his best-evidence synthesis approach resulted in such contrary findings to prior syntheses, Slavin compares a study that examines the effects of mastery learning on student achievement using both a researcher-constructed outcome measure and a standardized, norm-referenced achievement test outcome measure. The study is a one-year math mastery learning intervention study for students in grades 1-6 in one mastery learning and one control school in the same city. Slavin finds that mastery learning students perform significantly better on the researcher-constructed measure than the control group (mean ES=0.64); whereas, there was only a small non-significant difference between mastery learning students and the matched control students on the standardized achievement test that varied according to subscale (ES Computation=0.17; Problem Solving=0.07; Concepts=-0.12).

These findings lead Slavin to suggest that the previous syntheses (Block & Burns, 1976) and meta-analyses (Guskey & Gates, 1986) of mastery learning research *overestimated* the effects of group-based mastery learning on K-12 student achievement outcomes for two main reasons. First, because prior reviews included studies that he did not include since they did not meet the 4-week duration requirement. Second, because prior reviews relied almost exclusively on researcher- or teacher-constructed outcome measures that "correspond more closely to the curriculum taught in the mastery learning classes than to that taught in control classes" (p. 180). Slavin argues researcher- or teacher-made tests are particularly problematic in studies of mastery learning because mastery learning focuses students and teachers on a narrow and well-defined set of educational objectives where content mastery is emphasized over content coverage. This disadvantages control group students unless measures that include both content coverage and mastery are used as the dependent variable. Also, Slavin says researcher-constructed tests may have a ceiling whereby additional

learning cannot be registered on the outcome measure; whereas, standardized tests are unlikely to have a ceiling.

Another claim that Slavin examined had to do with the ability of mastery learning to effectively increase student achievement of specific skills or concepts central to a course of study (what he calls the "curricular focus" claim). He reviewed nine studies under this claim, 3 of which were examined under the first claim. In general, Slavin argued that the nine studies support the curricular focus claim. That is, findings tend to suggest that the effects of group-based mastery learning on researcher- or teacher-made measures are positive and moderate (Median effect size 0.27). His effect sizes were weaker than in the Guskey and Gates (1986) meta-analysis because Slavin said he adjusted effect sizes based on differences in the treatment and control groups at baseline and Guskey and Gates made no adjustments. Slavin also did not calculate effect sizes by pooling the standard deviations of both the treatment and control groups, but used only the standard deviation of the control group. He said this was because mastery learning "often has the effect of reducing achievement standard deviations" (p. 185), which would then make effects appear stronger if pooled.

Overall, results from Slavin's best-evidence synthesis are significantly different than previous research reviews and from later research reviews on the effects of mastery learning on K-12 student achievement outcomes. The results of Slavin's synthesis suggest that group-based mastery learning may have positive effects on student achievement *if the outcome measure is a researcher-constructed, criterion-referenced test* and may have no effect on student achievement *if the outcome measure is a standardized, norm-referenced achievement test*. There was also some evidence that suggests lower achieving students and low-SES students tended to experience greater effects of mastery learning, as demonstrated on researcher-constructed tests.

Limitations and Implications of Slavin's (1987a) Best-Evidence Synthesis

Immediately following Slavin's (1987a) article in the *Review of Educational Research*, Anderson and Burns (1987) and Guskey (1987) respond to Slavin's best-evidence synthesis research design and Slavin's claim that the effects of mastery learning have been overestimated. Anderson, Burns, and Guskey are all authors of previous research reviews and were considered experts on mastery learning (Block & Anderson, 1975; Block & Burns, 1976; Guskey & Gates, 1986). Their main points of disagreement with Slavin center around two main issues: the inclusion/exclusion criteria and how effect sizes were calculated. Slavin then responds to their critiques with his own rebuttal (Slavin, 1987b)

Regarding the inclusion/exclusion criteria, the experts on mastery learning (L. W. Anderson & Burns, 1987; Guskey, 1987) contend that Slavin's 4-week study duration criteria was arbitrary, based on a misinformed understanding of the philosophy behind mastery learning, and artificially eliminated many studies of 3-week duration that showed positive effects. Slavin (1987b) responds both by defending his choice as well as re-examining the studies of 3-week duration and says that even if he had included 3-week duration studies, his findings would not have been altered.

Another point of contention surrounding the inclusion/exclusion criteria was Slavin's (1987a) argument that the reason why prior reviews had overestimated the effects of mastery learning on student achievement was the use of researcher-constructed outcome measures that disadvantage the control group who may have covered more content without the ability to demonstrate that knowledge since it wasn't measured on the outcome measure. Anderson, Burns, and Guskey do not argue against Slavin's findings regarding no difference between treatment and control students on standardized achievement tests, but question the validity of standardized, norm-referenced achievement tests for mastery learning experiments. For example, Guskey (1987) argues that standardized achievement tests may be biased in favor of the control group because "they were

free to cover a greater number and wider range of objectives than mastery learning groups" (p. 227). In either case, both Slavin (1987b) and Guskey (1987) use the evidence of no difference between treatment and control groups on standardized test performance to arrive at different conclusions. Guskey (1987) says: "students in mastery learning classes did just as well on broad-based standardized measures as students in control classes…[which provides] strong evidence…that coverage was *not* sacrificed for the sake of mastery" (p. 227). Slavin (1987b) says: "group-based mastery learning has no important effects on standardized tests of reading and mathematics" (p. 231).

Regarding how effect sizes were calculated, the controversy surrounds how Slavin calculated effect sizes for studies with a pretest-posttest design. If the treatment and control students differed on pretest measures, Slavin used differences in treatment and control group *gains* not differences in treatment and control group means. Guskey (1987) argues this procedure has not been used in any other research synthesis to date and "serves mainly to systematically reduce all calculated effect sizes" (p. 227). Slavin (1987b) responds by defending his adjustment for pretests differences and then re-calculates effect sizes as if he had not adjusted, claiming that the median effect size estimate reported in his synthesis would have been the same for the studies using standardized tests (ES=0.04) since those are the ones that had pretest-posttest designs.

Slavin's (1987a) best-evidence synthesis and subsequent critiques of his methods have at least three implications for my dissertation study. First, it is important to justify the choice of an outcome measure and specify how the outcome measure does not disadvantage one group (treatment or comparison) so as to bias results. For this reason, the outcome measure should be equally fair in registering student achievement on both the depth and breadth of the content domain (coverage and mastery). Second, it may be important to calculate effect sizes using the standard deviation of the comparison group to prevent overestimating the effect of treatment, especially as

narrowing the variability in student performance is a potential outcome of the treatment in my dissertation study. And third, I may also find non-significant treatment effects, which as we have seen can be interpreted in multiple ways—no effect can be interpreted positively to mean that treatment students performed just as well and/or no effect can be interpreted negatively to mean that treatment was somehow ineffective. It is important to consider different ways of framing results and for the researcher to try to remove personal biases (to the extent that is possible) from the way study findings are reported.

Review of Guskey and Pigott's (1988) Meta-Analysis

Guskey and Pigott's (1988) meta-analysis was published a year after Slavin's (1987a) best-evidence synthesis, although it never mentions Slavin's review. This may be because Guskey and Pigott wrote their meta-analysis at the same time at Slavin. In addition, research published by Kulik, Kulik, and Bangert-Drowns (1990) was also written around this same time since it was presented at the American Educational Research Association annual meeting in 1986 (Kulik, Kulik, & Bangert-Drowns, 1986), even though it was not published until four years later. For simplicity, I chose to review these research reviews based on the published date in a peer-reviewed journal article rather than when they first presented their research at a professional conference.

Guskey and Pigott's (1988) meta-analysis is a follow-up to Guskey and Gates's (1986) meta-analysis on the effects of group-based mastery learning on student achievement outcomes measured using mainly researcher- or teacher-constructed end-of-course examinations. The researchers examine the research in five areas (student achievement, student learning retention, time variables, student affect, and teacher variables), but I focus my review here as I did earlier on student achievement results.

The key differences between the two meta-analyses are that the 1998 analysis includes more studies (43 studies on student achievement compared to 25 in Guskey & Gates, 1986) because it

expands the inclusion/exclusion criteria. Guskey and Pigott (1988) include both published and unpublished studies and expand the age range from K-12 to also include postsecondary studies. Another key difference between the two meta-analyses is that Guskey and Piggot expand upon the analytic methods used to examine the wide variation in effect sizes reported in Guskey and Gates. For example, in both meta-analyses the researchers found considerable variation in effect sizes (0.02 to 1.7) so that they chose not to calculate a mean or median effect size across studies. In the earlier meta-analyses by Guskey and Gates, the researchers examine sources of variance in effect sizes for two factors: subject area and grade span. Guskey and Pigott provide more detailed analysis of variance by subject area and grade span using other analytic techniques, and also analyze the variability in effect sizes by study duration as a third factor.

Guskey and Pigott (1988) analyzed 43 studies with a total of 78 effect size estimates. They calculated effect sizes as the difference between the mean scores of the two groups divided by the standard deviation of the control group. Almost all effect sizes were positive, meaning that mastery learning tended to have a positive effect on student achievement across studies. However, because the effect sizes varied so much study-to-study, the researchers tested for homogeneity of variance between effect sizes using a homogeneity statistic (H) that is based on a chi-square distribution[5] and found that the variation was "much greater than would occur if all studies shared an underlying effect size" (p. 203)($H=759.50$, $df=77$, $p<.001$). As stated earlier, the researchers tried to explain variation in effect sizes through examining differences for three factors (subject area, grade span, and study duration). They examined both the within-group and between-group variance for each factor, but did not examine two-way or three-way interactions among the factors.

In terms of the subject area analysis, Guskey and Pigott found that differences in effect sizes did vary both within subject area and between subject areas ($H_w=631.77$, $df=70$, $p<.001$; $H_b=127.73$,

[5] See Guskey & Pigott, 1988, pp. 203-204 for more information on the H statistic.

df=4, p<.001). They found that the greatest effects were exhibited in math followed by language arts, social studies, science, and psychology.

When examining variation in effect sizes by grade level the researchers also found that differences in effect sizes did vary both within grade spans and between grade spans (H_w=631.77, df=75, p<.001; H_b=127.73, df=2, p<.001). The greatest effects were seen in elementary/middle school (grades 1-8) followed by high school and then college. The researchers did not examine elementary and middle school separately, so it is unclear how elementary and middle school findings may have differed if examined separately. The effects in elementary/middle school were almost double the effects for high school.

Guskey and Pigott (1988) found that effect sizes did vary significantly within study duration categories (H_w=751.42, df=75, p<.001). Although the greatest effects on student achievement were for studies of 1-week duration followed by 2-12 weeks and then 18+ weeks, they did not find differences in effects between study durations to be statistically significant (H_b=2.09, df=2, p>.05). It is unclear why researchers chunked the study durations into those three categories or how findings may have varied based on different groupings. According to the researchers, no well-designed longitudinal studies were available to provide additional insight and these results run counter to Bloom's (1968) theory that mastery learning program effects would accumulate over time.

Overall, findings from Guskey and Pigott (1988) continue to support the positive effect of group-based mastery learning on K-12 student achievement if that achievement is measured using a researcher- or –teacher-made end-of-course test. This meta-analysis also provides more evidence about the variability in effect sizes from study-to-study and some potential sources of that variation (subject area and grade level). In sum, there is evidence that mastery learning tends to have the

greatest effects on student achievement in math and language arts at the elementary/middle school level on researcher- or teacher-constructed tests.

Limitations and Implication of Guskey and Pigott's (1988) Meta-Analysis

The limitations of Guskey and Pigott's (1988) meta-analysis are similar to those of earlier meta-analyses on the effects of group-based mastery learning on student achievement outcomes (Block & Burns, 1976; Guskey & Gates, 1986) therefore the limitations are briefly commented on. Obviously the inclusion of studies in this meta-analysis that mainly rely on researcher-constructed outcome measures is a limitation, especially given Slavin's (1987a) divergent findings. That said, there are a few other limitations to this particular meta-analysis worth noting: the use of several effect sizes from one study and how that violates assumptions of independence of observations; limited sources of variance in effect sizes analyzed; and lack of clarity for certain design decisions in the sources of variance actually analyzed.

First, Guskey and Pigott (1988) included 43 studies in their meta-analysis on student achievement, but used those 43 studies to calculate 78 effect sizes. However, calculating multiple effect sizes from a single study often based on the same groups, program, and setting would seem to violate the assumption of independence necessary for many statistical tests (including the H statistic that is based on a chi-square distribution). As Kulik et al. (1990) points out, this design decision "would also give undue weight to studies with multiple groups and multiple scales" (p. 270). A second limitation of Guskey and Pigott's meta-analysis is that they analyze sources of variance in effect sizes between- and within-studies using only three factors. It may have been helpful had they also analyzed how effect sizes varied as a function of the way mastery learning was defined in the studies or based upon the experimental designs of studies, for example. And finally, regarding the three sources of variation they do examine, Guskey and Pigott do not provide a rationale for why they pooled elementary and middle school studies together or why they categorized study duration

into three categories that are not equivalent (1 week, 2-12 weeks, and 18+ weeks). Furthermore, it would have been helpful if they had explained why they did not test 2-way or 3-way interactions between factors.

There are a couple implications of Guskey and Pigott's (1988) meta-analysis for this dissertation study. One implication is that the most likely place to find effects of a group-based mastery learning approach to instruction, which is similar in many ways to the competency-based education approach as applied in New Hampshire currently, may be in grades 1-8 in math and language arts. This provides a rationale and justification for focusing on effects at one of those grade levels and in those subject areas. Another implication is how they calculated effect sizes—using the standard deviation of the control group as the denominator—which can inform my own practices in this dissertation.

Review of Kulik, Kulik, and Bangert-Drowns' (1990) Meta-Analysis

Kulik, Kulik, and Bangert-Drowns' (1990) meta-analysis is a significant advance over prior meta-analyses on mastery learning in several ways. First, it is the largest review of the literature (N=108 studies). Their goal was to include the research studies reviewed in prior meta-analyses as long as they met four inclusion criteria: 1) studies had to be field evaluations of mastery programs; 2) students in the mastery programs had to be held to at least a 70% correct criterion for mastery; 3) studies needed to be free from serious methodological flaws; and 4) enough quantitative results needed to be reported so that effect sizes could be calculated or estimated. Another reason this study is an advance over previous meta-analyses is the way they analyzed the effect sizes. The researchers created fifteen variables based on information in the 108 studies to describe treatments, methodologies, settings, and public histories of the studies.

Most of the 108 studies used a researcher- or teacher-constructed, criterion-referenced examination as the outcome measure (N=103). Similar to Slavin (1987), Kulik and colleagues (1990)

calculated effect sizes by adjusting for pretest differences between the groups and dividing those gains by the standard deviation of the control group. They used only one effect size per study. Out of the 108 studies, 72 used Keller's Personalized System of Instruction approach at the college level and 36 used Bloom's Learning for Mastery group-based approach. Nineteen of the 36 group-based mastery learning studies took place at the college level and 17 took place at the K-12 level, although the primary grades K-2 were rarely included in studies. Interestingly, the 17 studies included in Slavin's (1987a) best-evidence synthesis of group-based Learning for Mastery elementary and secondary studies were not the exact same 17 studies included in this review (6 of the studies were different).

Kulik and colleagues (1990) had several important findings. First, 96 out of the 103 studies with student achievement data reported positive effects of mastery learning on student achievement (as measured on an end of unit exam). In 67 of the 96 studies the researchers found that the positive effects were statistically significant. The average effect size of the 103 studies was 0.52 standard deviations ($t=15.78$, $df=102$, $p<.001$), but effects did vary considerably from study to study. For example, the effect sizes ranged from 0.22 to 1.58.

To examine whether the variation in effect sizes might be systematic based upon different study characteristics or design features, the researchers used one-way analysis of variance with each of the 15 variables as factors. They found four study features were related to the size of the estimated treatment effect, including use of locally developed end of unit exam versus standardized achievement test and study duration. Similar to Slavin (1987a), the researchers found that the effect of mastery learning was very small on standardized outcome measures ($ES=0.08$).

The researchers then used multiple regression analysis to examine the conditional effect for each of the 15 variables from the studies. They found five variables had statistically significant effects on the prediction of effect sizes and explained about 25% of the total variance in effect

sizes—pacing (individualized or group)(B=0.14, p<.05); unit mastery level (B=0.01, p<.05); locally constructed vs. standardized test (B=-0.13, p<.01); amount of feedback for the control group (B=-0.21, p<.05); and subject matter (B=0.16, p<.05). To put those parameter estimates in context, group-based programs with high mastery standards were more effective. Also, programs whose effects were measured using locally constructed end of unit exams and who provided less feedback to the control group also had larger effects. Interestingly, mastery learning programs in the social sciences were more effective than programs in math and science. It does not appear that the researchers tested for any interactions.

Overall, Kulik and colleagues (1990) found that mastery learning programs had a small positive effect on K-20 student achievement as measured on a standardized achievement test (ES=0.08) and a moderate positive effect on locally constructed end of unit exams (ES=0.50), holding all else constant. The moderate effect was slightly larger when just Bloom's group-based Learning for Mastery programs were included (ES=0.59). Results also suggested that the effects of mastery learning programs may be stronger for students with lower prior achievement as measured on researcher-constructed pretests. For example, the average improvement for students with lower prior achievement was 0.61 standard deviations; whereas, the average improvement for students with higher prior achievement was 0.40 standard deviations. This is an important finding and it would be interesting to see if this finding holds on standardized prior achievement measures.

Limitations and Implications of Kulik, Kulik, and Bangert-Drowns' (1990) Meta-Analysis

Similar to the issue of *Review of Educational Research* three years prior, Slavin (1990) is given an opportunity to respond to Kulik and colleagues' (1990) meta-analysis. Slavin's main point of contention is with Kulik and colleagues' discussion on the value of researcher-constructed outcome measures. Slavin argues that the problem with including studies that rely exclusively on researcher-constructed exams as outcome measures is that the meta-analysis is then dominated by those

findings, which creates a biased overall effect size. Slavin also argues that the "substantial difference in outcomes on some of the studies that used both standardized and curriculum-specific measures demands some explanation" (p. 301). Kulik and colleagues respond by questioning some of Slavin's study selections and re-asserting their common ground.

The findings of Kulik and colleagues' (1990) meta-analysis differ from prior findings. For example, Guskey and Gates (1986) report stronger effects (mean effect size 0.78) than in this meta-analysis (mean effect size 0.52) on locally constructed tests. On the other hand, Slavin (1987) reports very weak effects (median effect size 0.25), which is much smaller than this meta-analysis (median effect size 0.43 for K-12 group-based mastery learning studies) on locally constructed tests. That said, Kulik and colleagues agree with Slavin that the average effect of group-based mastery learning on K-12 student achievement outcomes as measured on standardized tests is trivial (ES=0.08), but statistically significant ($t=3.0$, $df=4$, $p<.05$). This is slightly different from Slavin's findings where he did not find any significant effect of mastery learning on standardized achievement tests (ES=0.09, $t=2.3$, $df=4$, $p<.10$). Findings differ based on the inclusion/exclusion criteria applied in each research review.

Kulik and colleagues (1990) agree with Slavin's (1990) overall finding that it is important to use standardized achievement tests as outcome measures, although they also argue there is also value in locally constructed tests. They think both should be used as outcome measures. Kulik and colleagues also agree with Slavin that there are major differences in mastery learning effects depending upon whether student achievement is measured using locally developed or standardized tests. However, Kulik and colleagues would disagree with Slavin's interpretation of the meaning in those effect size differences and argue that locally developed exams may be more instructionally sensitive and therefore pick up the effects of an instructional intervention more readily than a standardized test.

There are a couple implications of Kulik, Kulik, and Bangert-Drowns' (1990) meta-analysis for my dissertation. First, it is important to examine treatment effects by subject area, student prior achievement, length of treatment, and the interactions among those variables. In this meta-analysis, they found differences in student achievement between subject areas (greater effect in social sciences) and level of student prior achievement (greater effect on lower prior achieving students). While the researchers did find that effect sizes varied based on study duration using one-way analysis of variance, study duration was non-significant when included as a predictor in the multiple regression analysis. Second, the continued debate in this meta-analysis over the validity of locally constructed tests as outcome measures for control students suggests that a criterion-referenced standardized achievement test may be equally valid for both treatment and comparison students and serve as a useful outcome measure for my dissertation study.

Review of Anderson's (1994) Meta-Synthesis

Anderson (1994) synthesized findings from six different meta-analyses, five of which are the same as what I reviewed above. He also included a meta-analysis by Willent, Yamashita, and Anderson (1983) that focused exclusively on science. I did not include that meta-analysis because it did not meet my inclusion criteria of relating to either English language arts or mathematics; however, I did include the meta-synthesis by Cotton and Savard (1982) that Anderson (1994) did not include. Overall, Anderson found that 224 of the 279 research studies included in the six meta-analyses (or 90%) show a moderate positive effect of mastery learning on student achievement in all subject areas and all grade levels (based on researcher-constructed end of unit exams), ranging from 0.27 to 0.94 standard deviations. Anderson also noted the evidence in some studies suggesting that mastery learning has greater effects in elementary school that decreases in size as students get older (Block & Burns, 1976; Guskey & Pigott, 1988). This may result from greater variability in student achievement, particularly as students get older.

Limitations and Implications of Anderson's (1994) Meta-Synthesis

Anderson's (1994) meta-synthesis draws together the major meta-analyses on the effects of mastery learning on K-20 student achievement outcomes. Anderson's synthesis is accurate, but general. It appears his purpose was to provide a broad overview of the effectiveness of mastery learning rather than drill down into the details of each meta-analysis. He does not trace the development of themes in the meta-analyses over time, expansion of analytic methods, or contradictions in findings. This is a limitation of this meta-synthesis and complicates efforts to draw implications from Anderson's study to this dissertation.

Synthesis Across All Mastery Learning Research Reviews

Rather than repeat the detailed analysis provided above, this synthesis across the seven research reviews on mastery learning begins where Anderson (1994) left off. Specifically, I trace the development of particular themes in the research reviews and expansion of analytic methods over time, as well as conclusions left open to debate. In so doing, particular attention is paid to what we know and what we have left to understand about the effects of mastery learning on K-12 student achievement outcomes in English language arts and mathematics. I interweave this section with the main implications of these studies for this dissertation study as they arise.

There are at least five themes that can be traced through the meta-analyses over time. First, *effect sizes vary considerably from study to study*. In some cases, this led the researchers to not calculate an average effect size across the studies included in the meta-analyses (Guskey and Gates, 1986; Guskey and Pigott, 1988). In other cases, the researchers did calculate an overall effect size (Kulik et al., 1990), but then used follow-up analysis to try to understand what features of mastery learning programs or study designs may be related to variation in effect sizes. Second, *overall effect size estimates weakened over time*. For example, effect sizes were reported as strong and positive in the earliest meta-

analyses (ES=0.83)(Block & Burns, 1976), but then were reported as more moderate to small[6] while still positive in later meta-analyses (ES=0.36-0.45)(C. L. Kulik et al., 1990; J. A. Kulik et al., 1990; Slavin, 1987a). A third theme was that *lower achieving students and/or low SES students may benefit more from mastery learning programs***.** Four meta-analyses discussed evidence that suggests lower achieving students may benefit more from mastery learning programs than higher achieving students (Cotton & Savard, 1982; Guskey & Gates, 1986; Kulik et al., 1990; Slavin, 1987). One meta-analysis also suggests that lower SES students may benefit more academically from mastery learning programs in comparison to high SES students (Slavin, 1987). These findings signal the importance of examining the extent to which treatment effects vary according to prior achievement and SES in this dissertation. A fourth theme was that *elementary and middle school students may show greater effects of mastery learning***.** For example, two similar meta-analyses found greater effects on elementary/middle school students in comparison to high school and college students (Guskey & Gates, 1986; Guskey & Pigott, 1988). And final theme is that *mastery learning tended to be associated with positive effects in all tested subject areas.* For example, most studies showed positive effects of mastery learning across subject areas. In terms of what subject area demonstrated the greatest effects of mastery learning on student achievement, the findings were inconclusive. Guskey & Gates (1986) found greater effects in social studies and language arts in comparison to math and science; whereas, Guskey & Pigott (1988) found the greatest effects in math.

Over time there were two main developments in terms of analytic methods across the research reviews on mastery learning. First, *meta-analyses became more detailed over time, eventually testing for variation in effect sizes along many different dimensions related to treatment, study design, and outcome measure.* The earlier meta-analyses provided global overviews of mastery learning effectiveness and only analyzed

[6] Slavin (1987a) reported a median effect size estimate of 0.27 on locally constructed tests and 0.04 on standardized achievement tests. J.A. Kulik et al. (1990) re-estimated Slavin's effect sizes using only the 11 studies that overlapped between their meta-analyses and reported mean effect sizes between 0.36-0.45 on locally constructed tests and 0.08-0.09 on standardized achievement tests.

study outcomes by type of mastery learning program (e.g., Keller's Personalized System of Instruction vs. Bloom's Learning for Mastery)(Block & Burns, 1976; Cotton & Savard, 1982). However, as time goes on, researchers began to examine the variation in effect sizes from study to study by other factors. The first two factors were subject area and grade level (Guskey & Gates, 1986). Slavin (1987) adds study duration as a factor alongside subject area and grade level, which Guskey and Pigott (1988) continue. Study duration was not related to effect size variability once other features of treatment and study designs were controlled (Kulik et al., 1990). The last meta-analysis by Kulik and colleagues (1990), however, greatly expands the number of dimensions used to explain variability in effect sizes as they test 15 different variables. Kulik and colleagues also expand the analytic methods used beyond one-way analysis of variance or other tests of homogeneity of variance to include multiple regression analysis.

A second development in analytic methods was that *effect size calculations became more precise over time*. The earlier meta-analyses did not adjust for pretest differences between treatment and control groups and pooled the standard deviations of the treatment and control groups (Block & Burns, 1976; Guskey & Gates, 1986; Guskey & Pigott, 1988). Later meta-analyses corrected for prior achievement differences between treatment and control groups using the difference in gains divided by the standard deviation of the control group to calculate effect sizes (Kulik et al., 1990; Slavin, 1987). This signals the need in this dissertation to calculate treatment effects using the comparison group's standard deviation (if it is larger than the pooled standard deviation).

There are two main disagreement that did not appear to be settled in the research on mastery learning: 1) the extent to which locally constructed outcome measures should be used in mastery learning studies and 2) how to interpret the difference in findings between locally constructed end of course examinations and standardized achievement tests. Overall, it was clear that the *effects of mastery learning on student achievement varied as a function of the type of outcome measure used in the study*. This is a

critical finding. Most studies on the effects of mastery learning used locally constructed outcome measures. A few used both locally constructed and standardized achievement tests. Slavin (1987) was the first to examine effects based upon the type of outcome measure used. Slavin found that effects of mastery learning were small, but positive on locally constructed exams, but effects were trivial on standardized achievement tests and not significant. Researchers after Slavin had to address his claim that the effects of mastery learning on student achievement had been overstated because the researchers relied on locally constructed outcome measures that were differentially valid for treatment students and not for control students (Anderson & Burns, 1987; Guskey, 1987; C. L. Kulik et al., 1990; J. A. Kulik et al., 1990). Others argued standardized (norm-referenced) achievement tests were not valid measures of mastery learning effectiveness because the alignment between the curriculum taught and what the test purports to measure were unclear (Anderson & Burns, 1987; Guskey, 1987). In addition, others argued that both outcome measures provide insight into the effectiveness of mastery learning programs because the locally constructed measure provides evidence about the effectiveness of mastery learning as an instructional intervention while the standardized test provides evidence about the effectiveness of mastery learning in balancing content mastery with content coverage (C. L. Kulik et al., 1990; J. A. Kulik et al., 1990). It is unknown how findings would differ if criterion-referenced standardized achievement tests had been used to measure student achievement. This signals the importance of choosing an outcome measure for this dissertation study that does not disadvantage either the treatment or comparison group because of the way treatment is implemented.

Competency-Based Education Studies

I turn now to review studies related to the recent resurgence of K-12 competency-based education in the United States. There are three studies and they are reviewed chronologically.

Review of Haystead (2010)

Haystead (2010) reports on a study where he compared seven school districts that employed the RISC (Re-Inventing Schools Coalition) model and eight non-RISC districts. The RISC model is similar to a competency-based model because key features include flexible pacing, personalized learning, and demonstration of proficiency upon readiness. He used the percentages of students who scored proficient or above on state tests in 2009 for reading, writing, and mathematics as the outcome measure. Haystead compared RISC schools to non-RISC school with similar demographic profiles based on urban/rural, ethnicity, and size of student populations within each of three states: Alaska, Colorado, and Florida. Haystead does not provide any descriptive statistics for the RISC vs. non-RISC schools or districts and there is no explanation of how schools and/or districts were matched or stratified on the three characteristics within each state. The outcome variable was a dichotomous variable at the student-level (proficient or above vs. below proficient), but school is the unit of analysis in this study so the outcome variable is the aggregate percentage of students scoring proficient or above. School-level data was aggregated for RISC and non-RISC schools to make the comparison. Haystead also included a researcher-created measure of RISC implementation level (low, medium, and high). To analyze the data, Haystead employed 2 x 2 contingency tables (RISC vs. non-RISC; Proficient vs. Not Proficient), odds and risk ratios, and the phi correlation coefficient. Approximately 3,900 students for each subject area were included although there is unexplained missing data in writing. It is unclear how many schools were included.

Haystead (2010) found that there were small positive associations between RISC schools and reading, writing, and mathematics proficiency rates around the magnitude of 0.2 ($p<.001$). Also, the

odds of a student scoring proficient or above at a RISC school are around 2.5 times larger on state tests in all subject areas than the odds of a student scoring proficient or above at non-RISC schools (reading=2.339; writing=2.503; math=2.433). Schools were also compared based on their level of RISC implementation. Findings from those analyses suggest that high implementing RISC schools have students who are more likely to score proficient or above on state tests in all subject areas in comparison to non-RISC schools or medium implementing RISC schools.

Limitations and Implications of Haystead (2010)

Findings from this study are limited for a number of reasons. First, the researcher did not account for the nested structure of the data especially that students, schools, and districts are nested within different states using different tests and different definitions of proficient. As shown in some of the Haystead (2010) results, the overall proficiency rates in Florida seem to be much higher in Florida than in the other states in at least one content area, making it more difficult to find differences in percent proficient between RISC and non-RISC schools. And given that we do not know anything about the distribution of elementary, middle, and high school students across the three states, it makes it difficult to interpret those results as well. In addition, the aggregation of data into a binary outcome variable (proficient or above/not-proficient) significantly reduces the variability in the data that can be explained or predicted by treatment. Also, the lack of detail and explanation about how comparison schools were chosen makes it unclear the extent to which the researcher is able to control for selection bias.

The implications of Haystead's (2010) study for this dissertation study stem directly from the limitations of Haystead's study design. For example, Haystead did not adequately address selection bias. As a result, schools implementing RISC may be systematically different from schools that do not implement RISC that also relates to the outcome variable, which biases study findings. This signals the importance of attempting to create equivalent treatment and comparison students at

baseline based on multiple dimensions/characteristics that are likely related to selection and outcome using propensity score methods. This dissertation study improves and expands upon Haystead's study methods because it is controls for student- and school-level observed characteristics that are potentially related to outcome in multi-level model specifications.

Review of RAND Study on Personalized Learning Schools (Bill & Melinda Gates Foundation, 2014; Pane et al., 2015)

These two research reports were included in this literature review even though the title refers to "personalized learning" because the concept of personalized learning as defined in these reports include: "learner profiles that enable each student to be known well; the development of personalized learning plans for students; progress based on demonstrated knowledge and skills, rather than seat time; and flexible learning environment" (Bill & Melinda Gates Foundation, 2014, p. 2). This definition well represents the competency-based model of flexible pacing, personalized learning, and progress based on demonstrated proficiency.

The two research reports are the same study except the 2015 report ("Continued Progress") includes three years of data (2012-2015); whereas, the 2014 report ("Early Progress") includes only two years (2012-2014). The study is on-going. All of the schools in the personalized learning study received funding from the Bill & Melinda Gates Foundation to implement personalized learning practices as defined above. However, each participating school could design and implement their personalized learning approach as desired. Schools are mostly located in urban areas with large populations of minority students from low-income families. There was no pre-intervention period for the personalized learning schools in these reports because each school was newly founded as a personalized learning school. Also, most of the schools are charter schools. In fact, in the 2014 report, all of the personalized learning schools are charter schools. The key requirement for inclusion in this RAND study is that schools had to receive funding from the Bill & Melinda Gates Foundation, have been implementing personalized learning practices for at least two years, and have

two years of assessment data. The researchers investigated common elements of personalized learning shared across schools, student achievement outcomes, and teacher/student perceptions. This review focuses on the student achievement outcomes analysis.

The main research question related to student achievement in this RAND study is: Did students attending the personalized learning (PL) schools make greater gains in math and reading over two or three years in comparison to a virtually matched comparison group? Similar to this dissertation study, the researchers attempted to address selection bias by comparing PL student performance to demographically similar non-PL student performance. To do so, however, the authors did not use propensity score methods, but instead used a virtual comparison group approach. In this approach, each PL treatment student can be matched with up to 51 students from NWEA's national testing database. Students were matched exactly on two criteria: the urbanicity of their school (urban, rural, suburban) and grade level. Students were also "approximately matched" (± 5 points on NWEA's Rasch Unit scale) based on a pretest MAP assessment. NWEA's testing database does not contain any other student-level covariates such as race/ethnicity, free and reduced price lunch status, disability status, or Limited English proficient status. One school-level "approximate matching criteria" was used—schools could not differ by more than 15 percentage points on the proportion of students who qualify for free- and reduced-price lunch.

The sample size for the 2014 report was 23 PL charter schools that served about 5,000 students and implemented from 2012-2014. The 2015 report includes 62 PL schools (57 charter and 5 district schools) that served approximately 11,000 students and implemented from 2013-2015, as well as continued following the initial 23 PL charter schools for another year (2012-2015). It is unclear how many non-PL schools and students were included in either report. The authors do report the covariate balance in the 2015 report technical appendix on three variables: student score

on the pre-test MAP assessment, school percentage of students eligible for free- or reduced-price lunch, and the elapsed time between pretest and post-test.

The RAND researchers analyzed the effect of attending a PL school on student achievement outcomes by comparing each PL student with his or her virtual comparison group of up to 51 students. The outcome measure was student growth on NWEA's (Northwest Education Association) Measures of Academic Progress (MAP) math and reading benchmark assessments administered to students on a computer each fall and spring. In other words, gain from pretest (fall) to post-test (spring) in the MAP assessment scale score was the outcome variable. The researchers state that they "fit statistical models that account for clustering of students within schools and of each student with his or her VCG of up to 51 students…[and that they] controlled for the percentage of students eligible for free or reduced-price lunch" (Pane et al., 2015, p. 39), but no other model specification details are provided. There are no tables provided with descriptive statistics on PL vs. non-PL students or schools. There are no tables provided with regression model parameter estimates or goodness-of-fit statistics. This makes it difficult to explore study findings in great depth.

The researchers report their study findings using effect sizes. In the 2014 report entitled "Early Progress" that includes the first two years (2012-2013 and 2013-2014 school years), the researchers found that students attending PL schools made gains in math and reading over the two years that were "significantly greater" than the virtual comparison group (p. 4). Gains translated into effect sizes of 0.41 in math and 0.29 in reading, pooling across all grades (K-12). When effects were disaggregated by grade span, the greatest gains occurred in the K-2 grade span followed by the 3-5 grade span. The weakest effects were observed in grades 6-8 (math=0.20 ES, N=884 students; reading=0.14 ES, N=934 students) and high school (math=0.22 ES, N=201; reading=0.14, N=289). This suggests that effects of competency-based approaches to education may vary as a function of

grade span with the weakest effects in middle school and that math effects are slightly stronger in magnitude than reading effects. The researchers did not analyze the data by grade level so it is unclear how effects vary within grade spans. Although results varied greatly among the 23 PL charter schools in the "Early Progress" study, the researchers report that two-thirds of PL schools had "statistically significant positive results in either math or reading" (p. 4). Furthermore, the researchers sorted students into quintiles based on baseline academic performance on the MAPS pre-assessment and found that PL students in every quintile, but especially the bottom quintile, had higher growth than their comparison students. This finding is purely descriptive, however, as no hypothesis testing was employed.

In the 2015 report, "Continued Progress," the schools that started implementing PL in 2012 were followed for an additional year (2012-2015) and another cohort of 62 schools (90% of which were charter schools) was examined that had only implemented for 2 years (2013-2015). Findings were similar to the earlier report except the effect sizes for the 62 schools are not as strong as for the 21 schools in the earlier cohort, pooling across all grades (K-12)—0.27 in math and 0.19 in reading in comparison to 0.41 in math and 0.29 in reading. Effects again tended to be larger in math than reading, as well as the elementary grades (K-5) in comparison to middle school (about 0.15 in math and reading), but different from the earlier cohort, the researchers found no difference between the PL students and non-PL students in math and reading performance in high school. Researchers also continued to see growth accumulate for students attending PL schools in the first cohort (i.e., third year of implementation). Due to there being no pre-intervention period, there was no way to examine the extent to which there were implementation dips.

Recognizing that study findings may be an artifact of selection bias resulting from the PL schools being mostly charter schools, the researchers used post hoc sensitivity analyses to examine

the robustness of study results. Researchers re-ran the VCG drawing only from other charter schools in the NWEA national database and they did not see differences in study findings.

Overall, findings from the two RAND studies of Bill & Melinda Gates Foundation funded PL start-up schools suggest that students attending these mostly charter, PL schools tend to exhibit greater academic progress over two years in math and reading on benchmark assessments in comparison to virtual comparison group peers. Results also suggest effects in math are stronger than in reading. Also, students who were the most behind academically made the most progress, which may allow them to catch up to their peers and relates to the purpose for the reform in the first place—closing the achievement gap.

Limitations and Implications of RAND Study on Personalized Learning

One limitation of RAND's study on personalized learning is the outcome measure chosen. NWEA's MAP is a benchmark assessment that has not been designed (or validated) for any accountability purpose and whose alignment with state curriculum frameworks has been questioned (Marion, 2011). There are only multiple-choice items on the MAP assessments—no constructed response or performance tasks; the assessment is therefore limited in the depth of knowledge it can measure (Marion, 2011). Furthermore, there is little evidence of the predictive validity of the MAP benchmark assessments to state achievement tests (Brown & Coughlin, 2007). Therefore, it is unclear if personalized learning would register a similar effect on student achievement using a state-level standardized achievement test. Would there be differential effects as seen in the research on mastery learning between two different outcome measures?

My dissertation improves upon this study design in the use of a state-level standardized achievement test that is designed to serve an accountability purpose (Smarter Balanced Assessment Consortium, 2015), whose alignment with the Common Core State Standards has been

independently examined (Doorey & Polikoff, 2016), and that contains different item types with four depth of knowledge levels (Herman & Linn, 2013).

Another limitation of RAND's study on PL is the threat of selection bias. Over 90% of schools included in the study were charter schools and all of the schools were founded specifically as personalized learning schools. Students attending these schools and the families who send their children to these schools are likely very different from other non-PL students. In order to create equivalent groups at baseline, unbiased estimates of treatment effects are predicated on the assumption that there are no unidentified or unobserved characteristics that predict assignment to treatment that are not included in the virtual comparison group model. However, there are many student- and school-level characteristics that are not included in the virtual comparison group matching procedure such as gender, race/ethnicity, IEP status, Limited English proficiency status, free and reduced price lunch status. My dissertation improves upon this study design in the use of propensity score methods that attempt to account for many observed pre-existing differences between treatment and comparison students related to selection.

Review of Steele and Colleagues' (2014) Study on Competency-Based Education

Steele and colleagues' (2014) study on competency-based education started around the same time as the studies on personalized learning just discussed. Both studies were conducted by researchers at RAND and both projects were funded by the Bill & Melinda Gates Foundation. I chose to review this study last because it is the most similar to this dissertation study in terms of analytic approach. This study helps to develop the analytic approach for this dissertation, particularly the methods Steele and colleagues used to examine effects in Philadelphia. To begin, I provide general background on the overall study and then review the research conducted in three sites (Adams 50, Asia Society, and Philadelphia) since the research designs and analytic methods varied in each site due to data constraints. I conclude this review by briefly synthesizing across all sites and

then discuss limitations and implications of the Steele and colleagues' (2014) study for this dissertation.

General Background. In 2011, the Bill & Melinda Gates Foundation created the Project Mastery grant program to support competency-based education initiatives in large school systems that serve a high proportion of disadvantaged youth. For example, the Project Mastery grants were awarded to generally large urban or suburban school districts in which more than half the students were minorities, although in most cases only a small percentage of the students in each district were exposed to the intervention. The grants took place during the 2011–2012 and 2012–2013 academic years and then ended. The first year was used for materials development and implementation took place in Year 2. The three recipient organizations carried out their pilot programs in a total of 12 public secondary schools distributed across five school districts in four states. Steele and colleagues' (2014) research includes an evaluation of the implementation, student experiences, and student performance for each of the three recipients. This review focuses on the research related to student achievement outcomes and is organized by recipient organization since the scope, treatment, implementation, timing, and analysis varied across sites. The main research question under review is: To what extent did students' exposure to competency-based education models predict their academic performance in mathematics or reading? The authors are careful to explain that their study is non-causal due to the research design and that all findings should be interpreted as descriptive in nature.

Adams 50. Adams County District 50 (Adams 50) is a large suburban school district in Colorado with about 10,000 students in 19 schools. In the 2008-2009 school year the district converted to a competency-based education system. However, the Project Mastery pilot (and therefore the "treatment" group in this study) was implemented by only seven teachers in grades 8 and 9 math—four teachers in the district's only high school and one teacher in each of the three

middle schools (N=551 students). The teachers created several math games and instructional videos, which was the extent of the implementation. Steele and colleagues (2014) describe Adams 50 as a "low-dose intervention" site for this reason. The researchers only examined math outcomes since the Project Mastery intervention was only math related.

The researchers did not conduct a school- or student-level analysis because only district-level data was publicly available. As a result, the researchers focused on analyzing district-level math performance 3 years prior to the competency-based intervention in 2008-2009 and four years after, which then doesn't provide any information on the Project Mastery "low-dose" intervention. They used a district-level synthetic comparison group (SCG) approach where other districts in the state were weighted according to their similarity to Adams 50 on a number of covariates. District-level covariates included baseline achievement, racial composition, free-and reduced price lunch status, and district size, although baseline achievement was given the greatest weight (0.912 out of 1.0).

Steele and colleagues (2014) found that there were "sizable" differences between Adam 50's district-level math performance in comparison to the SCG in the years following the competency-based education reform (p. 65). Adams 50 underperformed the SCG and the differences ranged from about -0.4 of a standardized math score in 2009 to -0.8 in 2013 (which is about 0.22 of a student SD). The researchers could not use traditional hypothesis testing because the SCG approach is non-parametric, but they did use a placebo test to argue that the magnitude of the effect was such that it was unlikely to have occurred by chance at the 10-percent level. The authors also found evidence for a large implementation dip in the year of implementation (2008-2009) and first year following the competency-based education reform (2009-2010) in Adams 50, but not in the SCG although there had been a downward trend for at least a few years prior in Adams 50. This was the reason why Adams 50 was interested in changing to a competency-based education system in the first place. In sum, Adams 50 underperformed the SCG in math based on what would have been

expected based mainly on baseline district math performance in the four years after the competency-based education reform; however, there was relatively no treatment and there are many confounding factors (such as selection bias) that could explain these results beyond the effects of the competency-based education reform.

Asia Society. Asia Society is a New York based nonprofit organization that partners with 34 schools in the United States. For the Project Mastery grant, Asia Society partnered with four public secondary schools that emphasized project-based learning and portfolio-based assessment: three of them urban charter schools in Denver and Houston and one of them a public, rural high school in Newfound, New Hampshire. Asia Society's Project Mastery pilot initiative focused on the creation of performance outcomes and rubrics at the 8th and 10th grade levels, sample curriculum modules, and professional development modules for teachers. About 1,064 students in the four secondary schools were exposed to the intervention. The researchers decided to focus their analysis on reading performance because that is where the focus of the reform took place. Analysis differed for Newfound in comparison to the Denver and Houston schools because of the data available.

In Newfound, a school-level synthetic comparison group (SCG) was used with similar covariates as Adams 50, but at the school-level. However, in contrast to the Adams 50 analysis, prior achievement scores were not weighted as heavily (only 0.21 out of 1); percentage of students substantially below proficient was weighted the most heavily (0.58). Newfound trends in school-level 11th grade reading performance in the two years prior to its adoption of competency-based education reforms in the 2009-2010 school year and two years after were compared to the trends in reading performance for the SCG.

Findings suggest both Newfound and the SCG declined in comparison to the state average reading achievement in the 2009-2010 school year, but Newfound "markedly outperformed" its SCG in the next two school years (p. 70). Newfound also outperformed the state as a whole by the

2011-2012 school year by about 0.4 of a school-level SD. The authors argue this is a "modest but nontrivial positive effect" (p. 70). Using the same type of placebo test as with Adams 50 the authors find that the range of 11th grade reading scores is within the range of estimates that would be expected by chance alone and should be interpreted "with caution" (p. 71).

In the other three Asia Society sites, there was no pre-intervention period since all three high schools had implemented various aspects of competency-based education since their founding as charter schools. The researchers did not use a SCG approach to compare similar schools, but rather used a covariate adjustment approach with OLS regression. Analysis also varied because the researchers did not have access to school-level scale scores, but only school-level percentage proficient. For the Houston site, the researchers used school-level percentage of students meeting or exceeding standards averaged across content areas (language arts, math, science, and social studies) for grades 9-11 as the dependent variable; whereas, in the Denver sites (2 high schools), the researchers did not pool across content areas but focused on reading proficiency. OLS regression was used with a vector of school-level demographic characteristics (racial composition, school size, LEP status, FRL status), year, and treatment status dummy variable as predictors. Steele and colleagues (2014) found that the Houston high school had proficiency rates that exceeded those of the state by about 17.87 percentage points (SE=6.424; $p<.01$), controlling for the school-level demographics and year.

For the Denver sites, data for the two schools were pooled and similar to Houston the dependent variable was the percentage of students meeting or exceeding state reading proficiency standards for a given grade and year combination (grades 6-10). Different from Houston, the researchers used multi-level modeling to account for the grade-level nesting within schools although it is unclear how the researchers used multiple levels with only school-level data. Model fit indices such as chi-square difference tests, as well as AIC and BIC estimates were not provided so it is

difficult to compare models. They found that the two treatment schools did show a higher proficiency rate, on average, controlling for the other variables in the model, but the difference was not statistically significant at the .05-alpha level (B=8.88, SE=7.11, p>.05).

Overall, results of the Asia Society investigation are inconclusive. There are positive effects of the competency-based education reform in reading for the four treatment secondary schools; however, those positive effects are not statistically significant in three of the schools. The analyses performed were also limited by the data that was available. Selection bias is a sizable threat to the validity of these study findings and the nesting of students within schools is not accounted for in any of these study designs.

Philadelphia. The Philadelphia School District is a large, urban district that served approximately 154,000 students in 266 schools during the 2011-2012 school year. Philadelphia used the Project Mastery grant with 8 teachers who volunteered from six high schools (N=528 students). Two of the high schools were charter schools. The teachers focused on developing and implementing new units for grade 9 ELA and one writing unit for grade 10 ELA. The Philadelphia analytic approach is the most similar to this dissertation study as the researchers had access to student-level data and were clearly able to distinguish which students received treatment or no treatment. The authors argue that the "dosage was relatively high, and we might reasonably expect to see a difference in outcomes between pilot and non-pilot ninth-grade classrooms in terms of student achievement" (p.78). It is important to note that the dosage relates only to the new teacher-created units in grade 9-10 ELA and so this is the first site where the Project Mastery treatment is examined and not competency-based education. This then limits the generalizability of study findings for this dissertation. It is also important to note that ninth-graders did not take an end of year accountability test in 2013, so the researchers measured academic achievement using two midyear benchmark tests—one administered in November 2012 and the other in January 2013. The researchers did not

specify what ELA benchmark test was administered so it is unclear the extent to which the outcome measure sampled the breadth and depth of the content domain and therefore accurately reflects student achievement.

The researchers specified different models, but their preferred model used propensity score weighting and multi-level modeling to estimate the average effect of treatment on the treated (ATT) using the percentage of items answered correctly on two ELA benchmark tests as dependent variables. The researchers explored several propensity score specifications (including a hierarchical logistic regression model with student- and school-level variables), but did not achieve the best student-level covariate balance using that approach. As a result, the researchers had to choose between propensity score specifications and decided to use only student-level covariates in the propensity score weighting. It is important to note that teachers selected into the grant, but the propensity score covariates were at the student-level. These student-level covariates included lagged prior achievement in reading *and* math (even though the outcome measure was only reading) from 7th and 8th grade, grade level dummy variable for lagged prior achievement, gender, race/ethnicity dummy variables, indicators for gifted, disability, limited English proficiency, free and reduced lunch, size of grade 9 cohort, and over age for grade status. The researchers also attempted average treatment effect (ATE) weighting, but this approach also yielded poorer covariate balance than with the ATT weights and therefore decided to estimate the ATT not the ATE.

Based on the preferred model specification, the researchers estimated that Project Mastery students scored 2.6 percentage points lower than non-treatment/comparison students on the first ELA benchmark test in November (~3 months of dosage)(SE=1.12, $p<.05$); however, Project Mastery students scored 0.86 percentage points higher than demographically similar students on the second ELA benchmark test in January, although this effect was not statistically significant (~5 months of treatment)(SE=0.84, $p>.05$). This suggests there may be some evidence for a slight

implementation dip in the first few months of a similar Project Mastery treatment that may then have positive effects within the first year. Unfortunately, the Philadelphia study was not long enough to trace student performance trends beyond 5 months and therefore does not provide any evidence of how student performance may differ between treatment and non-treatment students over time.

Synthesis Across Project Mastery Sites. Synthesizing across the Project Mastery sites, there is no clear evidence on the effects of competency-based education models on secondary student achievement. Most analyses focused on high school students in either reading or math. Student performance varied across sites and it is unknown whether this variability is an artifact of research design limitations such as limitations due to available data, selection bias, and/or differences in treatment or implementation plans across sites (to name a few confounds).

In Adams 50, for example, the researchers only had access to district-level math data and found that Adams 50 underperformed its synthetic comparison group in math in the 4 years after the competency-based education reform. However, it is unclear whether the large implementation dip resulted from the competency-based reform or whether there were other factors associated with Adams 50's low math performance not accounted for in the synthetic group comparison.

In the Asia Society sites the researchers had access to school-level data in reading, but in three of the four high schools the outcome measure was percent proficient in reading and there was no pre-intervention data to track performance trends before-and-after treatment because the three charter schools were founded as competency-based education schools. Therefore the positive, but non-significant effects noted in reading for three out of the four Asia Society high schools are inconclusive.

And lastly, Philadelphia serves as an example of a high dose intervention for the Project Mastery grant, but doesn't provide a lot of information on the effects of competency-based education. The extent of treatment was 8 teachers from six high schools who created new units for

their 9th and 10th grade ELA classes. It is unclear how this relates to competency-based education so the positive, but non-significant findings after 5 months in reading do not provide clear evidence related to expected findings for this dissertation study.

Limitations and Implications of Steele and Colleagues' (2014) Study on Competency-Based Education

As in the case with all studies, there are limitations to the Steele et al. (2014) study of Bill & Melinda Gates Foundation's three Project Mastery grant awardees. One major limitation was that the researchers spent time explaining the Project Mastery intervention in each of the three sites, but then could not compare treatment to non-treatment in almost every site because of the way treatment was implemented (or not implemented) and the data they had available to analyze. For example, in Adams 50 (a district of 10,000 students and 19 schools) only 7 math teachers in two grade levels volunteered to participate and the extent of treatment was a few math games and instructional videos. If the researchers had access to student-level data they could have examined outcomes for students who had those 7 teachers for math in comparison to students in demographically similar classrooms from the same district, but unfortunately the researchers only had access to district-level data. But even if they did have student-level data the intervention was so "low dose" it is unclear what effects would even mean. Contrast this with the treatment in Philadelphia, which was considered "high dose" by the researchers, but only consisted of teacher-created units in 9th grade ELA and one writing unit in 10th grade ELA. No explanation is provided as to why these units are competency-based. The 8 teachers who participated also volunteered to participate in the pilot and students were not randomly assigned to classrooms so the extent to which the researchers were able to disentangle teacher and classroom effects from treatment effects is also unclear.

Given the limitations just discussed, as well as the limitations discussed within the general review and synthesis, there are at least two implications of Steele and colleagues' (2014) study that

can be applied to this dissertation study. First, the researchers use double robust estimation methods where possible. For example, in Philadelphia because the researchers have access to student-level data and can identify treatment from non-treatment status at that same level, the researchers use both propensity score and regression methods. This is considered a double robust estimator of treatment effects because only "1 of the 2 models need be correctly specified to obtain an unbiased effect estimator" (Funk et al., 2011, p. 761). The doubly robust analytic approach in this dissertation study was modeled after this example.

A second implication of the Steele and colleagues (2014) study is that the researchers used *a priori* criteria to justify the choice of propensity score model specification—even if I don't agree with their final choice. For example, the main criteria in deciding which covariates to include in the propensity score model should be the level of selection. However, the researchers did not use the classroom-level, but the student-level even though it was teachers who selected into the pilot not students. This implies that I should clearly delineate the criteria for including covariates in the propensity score model and connect it to the potential selection bias mechanism.

One justification for this dissertation study, provided in Steele and colleagues' (2014) six lessons for policy and practice, is that they call for the assessment of competency-based education programs on both near-term (such as the first few years of implementation) and longer-term outcomes using achievement test scores (see, for example, p. xvii). This dissertation responds to that call by examining two years of outcome data. This dissertation also extends the Steele et al. (2014) study because it estimates treatment effects in both ELA *and* math at the student-level and includes a significantly larger sample of schools and students. One difference between this dissertation and the Steele et al. study is that this study focuses on 8th grade students; whereas, their study included both middle school and high school students.

Synthesis Across Competency-Based Education Research

This synthesis focuses on what is known and what is left to understand about the effects of K-12 competency-based education on student achievement outcomes in math and ELA. I was able to locate 3 different studies that examine the effects of competency-based education on K-12 student achievement. These studies all took place in the last 10 years and each study examines schools in multiple states.

Most of the research in this area attempted to compare student performance on standardized tests for students receiving the competency-based treatment to those students not receiving any competency-based treatment. In most cases, the researchers only had access to school- or district-level aggregate data. Different methods were used to account for selection bias, including: stratifying (Haystead, 2010), synthetic comparison groups (Steele et al., 2014), virtual comparison groups (Bill & Melinda Gates Foundation, 2014; Pane et al., 2015), and propensity score weighting (Steele et al., 2014-Philadelphia). The most common characteristics accounted for in these methods were prior achievement, free- and reduced-price lunch status, race/ethnicity, and size of cohort or school/district. Steele and colleagues' (2014) study in Philadelphia arguably used the most robust method for controlling for selection bias. They used propensity score weighting and included nine student-level covariates: prior achievement, gender, race/ethnicity, size of cohort, and indicators for gifted, disability, limited English proficiency, free- and reduced-price lunch, and over age for grade status. This dissertation study builds upon the Steele et al. study, but uses covariates that match the level of selection.

Outcome measures varied study-to-study. In some studies, the percent proficient or above was used as the outcome measure (Haystead, 2010; Steele et al., 2014-Denver & Houston); whereas in other studies the researchers had access to scale scores on state achievement tests (Steele et al., 2014-Adams 50 & Newfound) or benchmark assessments (Bill & Melinda Gates Foundation, 2014;

Pane et al., 2015; Steele et al., 2014-Philadelphia). Using a binary outcome measure (percent proficient or above vs. not proficient) limits the explainable variability in outcome and is less preferable than using the entire scale score range. There are also limitations with the use of benchmark assessments as outcome measures. This dissertation extends the research in this area because it relies on student scale scores on a state achievement test as the outcome measure.

Researchers also used different types of analyses to estimate treatment effects. Researchers were sometimes limited in their analytic approach because of the available data. The most robust treatment effect estimates resulted from the use of regression analyses that took into account the nested structure of the data (students nested within schools)(see, for example, Steele et al., 2014-Philadelphia). This dissertation builds upon the strengths of prior research in this area by conducting a student-level analysis of treatment effects using multi-level modeling. This dissertation also expands upon the prior research in this area by examining differential effects for students based on their free- and reduced-price lunch status, disability status, gender, and prior achievement level.

Overall, findings from across the K-12 competency-based education studies were generally inconclusive. There is some evidence to suggest that there might be small positive effects on K-12 student achievement in reading and math in charter schools founded with competency-based education models after two years (Bill & Melinda Gates Foundation, 2014; Pane et al., 2015; Steele et al., 2014-Denver & Houston). There is not enough evidence yet to speculate about effects of competency-based education models on K-12 student achievement outcomes in public schools not founded as competency-based or personalized learning schools. There is also not enough evidence to understand the extent to which competency-based treatment relates to implementation dips or how long those implementation dips last. Similar to the research on mastery learning, there is some evidence to suggest that effects may be greater for elementary students than middle school and high school students and that the lowest performing students may benefit the most (Bill & Melinda Gates

Foundation, 2014; Pane et al., 2015). Based upon these findings and the limitations of studies to date, there is a need for further research on the effects of competency-based education on student achievement outcomes in all grade levels and all subject areas.

Summary of Prior Literature & Rationale for Study Design

In this section, I summarize across the two bodies of literature (performance assessment program research and mastery learning/competency-based education research) to draw out implications of the prior literature for this dissertation's research design and expected findings. Specifically, this section explains how the research methods in this dissertation study build on and address the strengths and limitations of prior research, as well as what the prior literature foreshadows in terms of the expected findings from this dissertation.

Research Design

There are many ways this dissertation builds upon the limitations and strengths of the prior literature. First, one of the limitations noted in the research on the effects of performance assessment programs on K-12 student outcomes was the difficulty extrapolating findings from the survey research designs. Researchers tended to examine relationships between teacher perceptions about the performance assessment program or their self-reported changes in instructional practices resulting from the reform with K-12 student achievement outcomes. However, most researchers had difficulty making sense of unusual or nonsensical results, which led some to question the sensitivity of their survey instruments (Stecher et al., 1998; Stecher & Chun, 2001). Other researchers found teachers may over-report their use of reform-oriented instructional practices, which further complicates relationships between reported practices and student achievement (Parke et al., 2006). This dissertation study improves and adapts to complications noted in the prior literature because it utilizes a different research design. As discussed earlier, this dissertation investigates the extent to which there is any effect of treatment on student achievement. Future

studies may then explore potential reasons for differences in effects between treatment and comparison groups based on contextual differences captured in surveys or artifact analysis such as fidelity of implementation and/or teacher perceptions of the reform alongside differences in achievement trends.

In contrast, one of the strengths of some of the research literature on the effects of competency-based education on K-12 student achievement outcomes is the research design. Specifically, some researchers used double robust estimation methods such as propensity scores and regression to maximize their ability to obtain unbiased treatment effect estimates (Steele et al., 2014-Philadelphia). This dissertation draws on the work of Steele and colleagues (2014) in this approach. This dissertation also draws on the prior literature to inform the choice of student- and school-level covariates that should be included as control variables in the propensity score and/or regression model such as: prior achievement, gender, race/ethnicity, size of cohort, disability status, limited English proficiency, and free- and reduced-price lunch. This dissertation extends the prior literature by also examining differential effects for certain subgroups of students.

Another limitation in some of the research on the effects of mastery learning and competency-based education on K-12 student achievement outcomes was the use of either a researcher-constructed or a benchmark assessment outcome measure. For example, researcher-constructed outcome measures may disadvantage the comparison group and benchmark assessments may not measure the full breadth and depth of the content domain. This dissertation responds to the potential limitations inherent in these outcome measures and relies on a third type of outcome measure also used in some of the prior literature—standardized achievement tests.

Another strength of some of the prior literature was the duration of treatment and inclusion of multiple years of data to examine effects over time. Some of the recent research on competency-based education included two years of data (Bill & Melinda Gates Foundation, 2014; Pane et al.,

2015) or examined trends in student performance before-and-after implementation with up to five years of data (Steele et al., 2014-Adams 50 & Newfound). Furthermore, the research on performance assessment programs suggest that one year of treatment may be too little to see evidence of treatment effects (Shepard et al., 1995), but treatment effects were evident at least after three years (Lane et al., 2002; Parke et al., 2006; Stone & Lane, 2003). This dissertation builds on this prior literature by examining treatment effects over two years with successive cohorts of implementers. This may allow some insight into dosage effects and implementation dips for this study population.

Expected Findings

In terms of expected findings from this dissertation, there are several factors that may complicate finding an effect of the PACE pilot after two years. For example, there are likely many reforms taking place within districts at the same time. Some of these reforms may work in concert with the theory of action behind the PACE pilot whereas others may not. Also, as in any new educational intervention/program, there is a learning curve. This learning curve may result in implementation dips and those implementation dips may last for multiple years (Fullan, 2001). Fidelity-of-implementation most likely also varies among the PACE schools/districts and is not accounted for in this dissertation study. Each of these factors may result in off-setting effects because one school may perform higher than expected whereas another school may perform lower than expected. This is one of the reasons why this dissertation examines the school-level residuals for each PACE school and not just average effects. And yet, some of the design features of the PACE assessment and accountability system described in detail in Chapter 3 under "Study Context" might help offset issues with duration and fidelity-of-implementation to some extent. In particular, the tiered system of rolling cohorts and additional implementation, professional development, and capacity building supports provided to participating districts.

That said, the use of a standardized outcome measure may also present some problems in finding a PACE effect if one exists in the population. For example, if part of the PACE theory-of-action is that standardized achievement tests have negative effects on curriculum and instruction such that curriculum loses its depth and instruction loses its complexity then it is possible that PACE districts spend less time on test preparation and place less emphasis and importance on standardized achievement tests results. If this is the case, improved performance exhibited by PACE students would have to be large enough to offset any reduction in test performance among PACE students because of lack of test preparation and/or enhanced (perhaps inflated) test performance among non-PACE comparison students in schools/districts that have focused on test preparation and test performance.

Given these potential confounds and considerations, there is some evidence from the prior research literature to suggest that there may be small positive effects of a performance assessment program in a competency-based learning environment on K-12 student achievement outcomes after two years. These claims are based on the research syntheses of each body of literature above where small positive effects were registered on standardized achievement tests in some cases (e.g., Anderson, 1994; Bill & Melinda Gates Foundation, 2014; Haystead, 2010; C. L. Kulik et al., 1990; Pane et al., 2015; Shepard et al., 1995; Slavin, 1987a; Steele et al., 2014; Stone & Lane, 2003). I could not locate any decisive evidence about whether effects tend to vary as a function of subject area (ELA or math) as findings tended to vary from study to study within bodies of literature. It may be the case that positive effects are easier to find in math early in the implementation of the PACE pilot, which would align with findings from the performance assessment research literature (Shepard et al., 1995; Stone & Lane, 2003).

That said, there is evidence from the competency-based research literature that effects may vary by grade level (Bill & Melinda Gates Foundation, 2014; Guskey & Gates, 1986; Guskey & Pigott, 1988; Pane et al., 2015). For example, in some studies elementary students tended to exhibit greater effects of mastery learning or competency-based education reforms in comparison to middle school and high school students. This mirrors findings from other educational intervention research syntheses where average annual gains in effect size from nationally normed tests are the largest in lower elementary grades and decline steadily into high school (C. J. Hill, Bloom, Black, & Lipsey, 2008).

Also, this dissertation draws on earlier research by examining the extent to which treatment effect estimates vary according to student prior achievement and as a function of student free- and reduced-lunch status. Prior research suggests that lower achieving and/or lower SES students tend to exhibit greater effects of mastery learning or competency-based education reforms in comparison to higher achieving and/or higher SES students. It is unclear whether this pattern holds in this population, especially as New Hampshire's lower SES students are not necessarily concentrated in urban areas. This dissertation also extends prior research by examining the extent to which treatment effects vary according to other observed student characteristics such as disability status and gender—neither of which were examined in the prior research literature in these areas.

Chapter 3: Study Design

The purpose of this dissertation is to examine the effects of an innovative assessment and accountability system. Chapter 1 explains why the research questions are being asked. Chapter 2 describes what we know and have yet to understand about the research question. Chapter 2 is grounded in the prior empirical literature on the effects of performance assessment programs and competency-based education on K-12 student achievement outcomes. In this chapter, I describe how the research question is going to be answered. I begin by explaining the context, which includes the history of competency-based education and performance-based assessment in New Hampshire, as well as providing a detailed overview and history of the treatment under investigation—New Hampshire's Performance Assessment of Competency Education (PACE) pilot program. I then provide a description of the datasets, population, sample, measures, and analytic approach used in the study. The design decisions are justified and the methods and procedures are described. This step-by-step description may assist future researchers who seek to replicate this study.

Study Context

New Hampshire (NH) is a small state located in the northeastern part of the United States. According to the U.S. Census Bureau (2015), New Hampshire is the 9th least populated of the 50 states, with an estimated population around 1.33 million people in 2015. Around 20% of NH's population is persons under 18 years old, which is similar to the percent for the entire United States (see Table 3.1)(U.S. Census Bureau, 2015). The majority of NH's population is White (93.9%) in comparison to the national percentage (77.1%) with the next largest race/ethnicity in NH being Hispanic or Latino (3.4%). Ninety-two percent of persons age 25 years old and up graduated from high school in NH, which is almost 6 percentage points more than the national average. The median household income in NH is also higher than the national average and the poverty rate is lower in NH than nationally. In fact, in a recent report the NH median income was the highest in the country

(Ingraham, 2017). It is worth noting that the New Hampshire context—though important in its own right—may not be representative of other states nationwide (a potential limitation explored in greater detail in the conclusion). That said, New Hampshire is also similar to many other states because it has a large rural population. Therefore, many of the same challenges faced by other rural states are seen in New Hampshire and faced by New Hampshire's educational system.

Table 3.1 Demographics for New Hampshire vs. United States

	New Hampshire	United States
Population	1,330,608	321,418,820
Persons under 18 years	19.8%	22.9%
Race/Ethnicity		
White	93.9%	77.1%
Black or African American	1.5%	13.3%
American Indian or Alaska Native	0.3%	1.2%
Asian	2.6%	5.6%
Native Hawaiian and Other Pacific Islander	0.0%	0.2%
Two or More Races	1.6%	2.6%
Hispanic or Latino	3.4%	17.6%
Education		
High school graduate or higher, percent of persons age 25 years+, 2010-2014	92.0%	86.3%
Bachelor's degree or higher, percent of persons age 25 years+, 2010-2014	34.4%	29.3%
Income and Poverty		
Median household income (in 2014 dollars), 2010-2014	$65,986	$53,482
Persons in poverty	8.2%	13.5%

Note. All statistics are estimates as of July 1, 2015 from the U.S. Census Bureau (2015) website. a=data from NH DOE website.

Although NH has historically been a top performer in the country with high graduation rates and standardized test scores, the state has continued to innovate its K-12 educational system (New Hampshire Department of Education, 2014a). Arguing from the vantage point of an outdated traditional learning model that valued time spent in the classroom instead of mastery learning, shifting workforce needs, and concerns with math and ELA proficiency rates decreasing over the course of a student's K-12 education, NH moved to a competency-based education system and performance-based assessment and accountability system in the last 10 years (New Hampshire

Department of Education, 2014a). Currently, NH is considered a leader nationally in terms of competency-based education and innovative assessment and accountability system reforms (Rothman & Marion, 2016). A brief history of these reform efforts is detailed below.

History of Competency-Based Education in New Hampshire

The notion of K-12 competency-based education began in New Hampshire in the 1990s with the school-to-work movement (M. Gfroerer, personal communication, November 21, 2016). In this movement, the New Hampshire Department of Education (NHDOE) developed a model transcript that included non-cognitive behavioral qualities or habits of mind alongside academic achievement indicators. Now called "work study practices," these behaviors are those that "students need to be successful in college, career, and life" (New Hampshire Department of Education, 2014c). These practices include, but are not limited to: listening and following directions, accepting responsibility, staying on task, completing work accurately, managing time wisely, showing initiative and being cooperative (NHDOE, 2014b). The NHDOE and its partners quickly realized that to ensure work-study practices were incorporated into K-12 education, teaching and learning practices (including assessment practices) needed to change.

As a result, in 1997, the NHDOE widened the project to include project-based learning, hands-on learning, and performance assessments (M. Gfroerer, personal communication, November 21, 2016). This project was known as the Competency-Based Assessment System (CBAS) and started with bringing teams of schools together to write competencies and then create assessments to measure student progress towards those competencies. The project began at the high school level, but then expanded to K-8 and also expanded from work-study practices to include academic competencies. According to Freeland (2014), CBAS included 30 schools by 2003. However, the state budget for CBAS was eliminated in 2003 for political reasons including the requirement to begin

implementing a statewide achievement test under NCLB, although schools continued to implement CBAS on their own (M. Gfroerer, personal communication, November 21, 2016).

Beginning in 2004, the state began convening key stakeholders to "redefine the goals and design of the state's high school system" (Freeland, 2014, p. 4). This led to a new vision for New Hampshire's high schools that focused on student-centered, personalized learning with real-world application (New Hampshire Department of Education, 2005). As part of this broad reform effort, New Hampshire was the first state to eliminate the Carnegie unit (or hours of class time required for every student to graduate) in 2005 with revisions to the "Regulation Education 306, the Minimum Standards for Public School Approval" (Ed 306) (Freeland, 2014). Ed 306 stipulated that all NH school boards require high school credit be earned by demonstrating sufficient mastery of required course competencies identified or developed by September 2008. However, each district was given "enormous latitude" to define competency-based education, decide on appropriate ways to assess competency, and define competency within their district (Freeland, 2014, p. 5).

The state attempted to provide assistance and guidance to districts beginning in 2013 through creating statewide college and career ready competencies in ELA, math, and science. The NH state model high school ELA and math competencies are aligned to the Common Core State Standards and were approved alongside high school science competencies by the State Board of Education for statewide use as of May 2014 (NHDOE, 2014c). The state has since expanded its efforts to include state model K-8 ELA, math, and science competencies, K-12 arts competencies, and work-study practices competencies (NHDOE, n.d.).

History of Performance-Based Assessment in New Hampshire

At the same time that competency-based education was developing in the state, there was also interest from the NHDOE in returning to a classroom-based assessment system that would incorporate performance assessments into the state assessment and accountability system (M. Gfroerer, personal communication, November 21, 2016). However, the regulatory requirements under NCLB that a state-level achievement test be administered every year to students in grades 3-8 and once in high school slowed the progress on a state-level performance assessment program. It wasn't until the 2012-2013 school year that the state contracted with the Center for Collaborative Education (CCE) to provide professional development to NH educators around performance assessments. This New Hampshire Performance Assessment Network involved cohorts of school districts (around 20 total) who expressed interested in assessment literacy training, as well as designing high-quality performance tasks and reliably scoring them using within-school protocols. CCE conducted three Quality Performance Assessment (QPA) training sessions over the course of the year.

In the next school year (2013-2014), four school districts that had participated in the QPA training and demonstrated progress and interest in "going deeper" during the training sessions around designing, administering, and scoring performance assessments were invited by the NHDOE to be help design a pilot statewide performance assessment accountability system (M. Gfroerer, personal communication, November 21, 2016). The National Center for the Improvement of Educational Assessment (NCIEA) was contracted at this time to provide technical expertise on the design of this innovative assessment and accountability system. These four districts, as well as the other districts involved in the NH Performance Assessment Network continued to participate in QPA training through CCE during the 2013-2014 school year.

The four school districts that were invited by the NHDOE to voluntarily implement PACE in Year 1 of the pilot (2014-2015 SY) were the same four school districts that helped to design the system in the previous school year after attending QPA trainings. An additional four school districts that had been part of the NH Performance Assessment Network and QPA trainings self-selected to participate in Year 2 of the pilot (2015-2016 SY)(see Table 3.2)

Table 3.2 List of school districts implementing the NH PACE pilot by year

Pilot Year	District ID Numbers
Pilot Year 1 (2014-2015)	Cohort 1: 165, 461, 476, 493 (high school only)
Pilot Year 2 (2015-2016)	Cohort 2: Year 1 Districts + 111, 365, 439, 705

Overview of the Performance Assessment of Competency Education (PACE) Pilot

In the fall of 2014, the NHDOE applied for a 2-year waiver (2014-2015 and 2015-2016 school years) from federal statutory requirements related to annual state-level achievement testing (NHDOE, 2016b). The U.S. Department of Education officially approved NH's Performance Assessment of Competency Education (PACE) pilot by granting a waiver in March 2015, allowing selected NH school districts to base annual determinations of student proficiency in ELA and math in grades 3-12 on a combination of local, common, and state-level assessments (Table 3.3)(NHDOE, 2014). The pilot was granted additional one-year waivers for the 2016-2017 and 2017-2018 school years.

Table 3.3 Local, common, and state-level assessments used to make annual determinations in NH's PACE pilot project

Grade	ELA	MATH
3	Smarter Balanced Achievement Test	Common and Local Assessments
4	Common and Local Assessments	Smarter Balanced Achievement Test
5	Common and Local Assessments	Common and Local Assessments
6	Common and Local Assessments	Common and Local Assessments
7	Common and Local Assessments	Common and Local Assessments
8	Smarter Balanced Achievement Test	Smarter Balanced Achievement Test
9	Common and Local Assessments	Common and Local Assessments
10	Common and Local Assessments	Common and Local Assessments
11	SAT	SAT

Local assessments include all summative assessments given within districts to assess student progress towards competency. Common assessments (i.e., PACE Common Tasks) are performance assessments created by representatives of all participating PACE districts and administered by all participating PACE districts in every grade and subject area where there is not a state-level achievement test. The common assessments or PACE Common Tasks are used to calibrate scoring across districts and enhance the comparability of annual determinations of student proficiency (see Evans & Lyons, 2017 for a detailed explanation). In the PACE pilot, state-level achievement testing occurs once per grade span. Annual determinations of student proficiency in PACE districts are based on common and local performance-based assessments alongside teacher judgment surveys except in those grades and subject areas where the state achievement test is administered. Figure 3.1 provides an example of a PACE Common Task (common performance assessment) for high school geometry. All PACE Common Tasks are scored using multi-dimensional analytic rubrics with 4-performance levels. There are inter-rater reliability audits that take place within districts and comparability audits that take place across districts (see the PACE Technical Manual for more details; Lyons, Evans, Marion, Thompson, 2017).

Figure 3.1 Example of a PACE performance assessment from high school geometry

- **The Problem:** Your town's population is predicted to increase over the next 3 years. As one of the town planners, you are asked to address this issue in terms of the town's water supply. In order to meet the future needs of the town, you need to make a proposal to add a water tower somewhere on town property that will be capable of holding 45,000 ± 2,000 cubic feet of water. The town is looking for a water tower to contain the most amount of water while using the least amount of construction material.
- **Student Task:** Your job is to prepare a proposal that can be submitted to the town planning committee. Using your calculations of surface area and volume for the two designs, describe and analyze the characteristics that lead you to a final recommendation.

High-quality performance assessments play a crucial role in the PACE system because of the need to measure the depth of student understanding on key competencies. Performance assessments are used both to inform teachers and students of how the learning activities are working and what might need to be adjusted (formative) along with serving to help document what students have learned (summative).

PACE Theory of Action

The PACE theory of action is grounded in the latest advances related to how students learn (Lave & Wenger, 1991; National Research Council, 2000; Shepard, 2000), how to assess what students know (National Research Council, 2001), and how to foster positive organizational learning and change (Elmore, 2004; Fullan, 2001; Pink, 2009). Figure 3.2 illustrates a version of the PACE theory of action with system design features on the left to outcomes on the right. The purpose of this theory of action is to illustrate broadly how implementation of the PACE system is intended to impact the instructional core of classroom practices (City, Elmore, Fiarman, & Teitel, 2009), thereby advancing college and career readiness. In its most basic form, the theory of action postulates that system design features drive changes to the instructional core of classroom practices such that teachers focus on the depth and breadth of key competencies (or content standards). These changes in instruction then lead to improved student achievement outcomes for all students; specifically, that students are college or career ready.

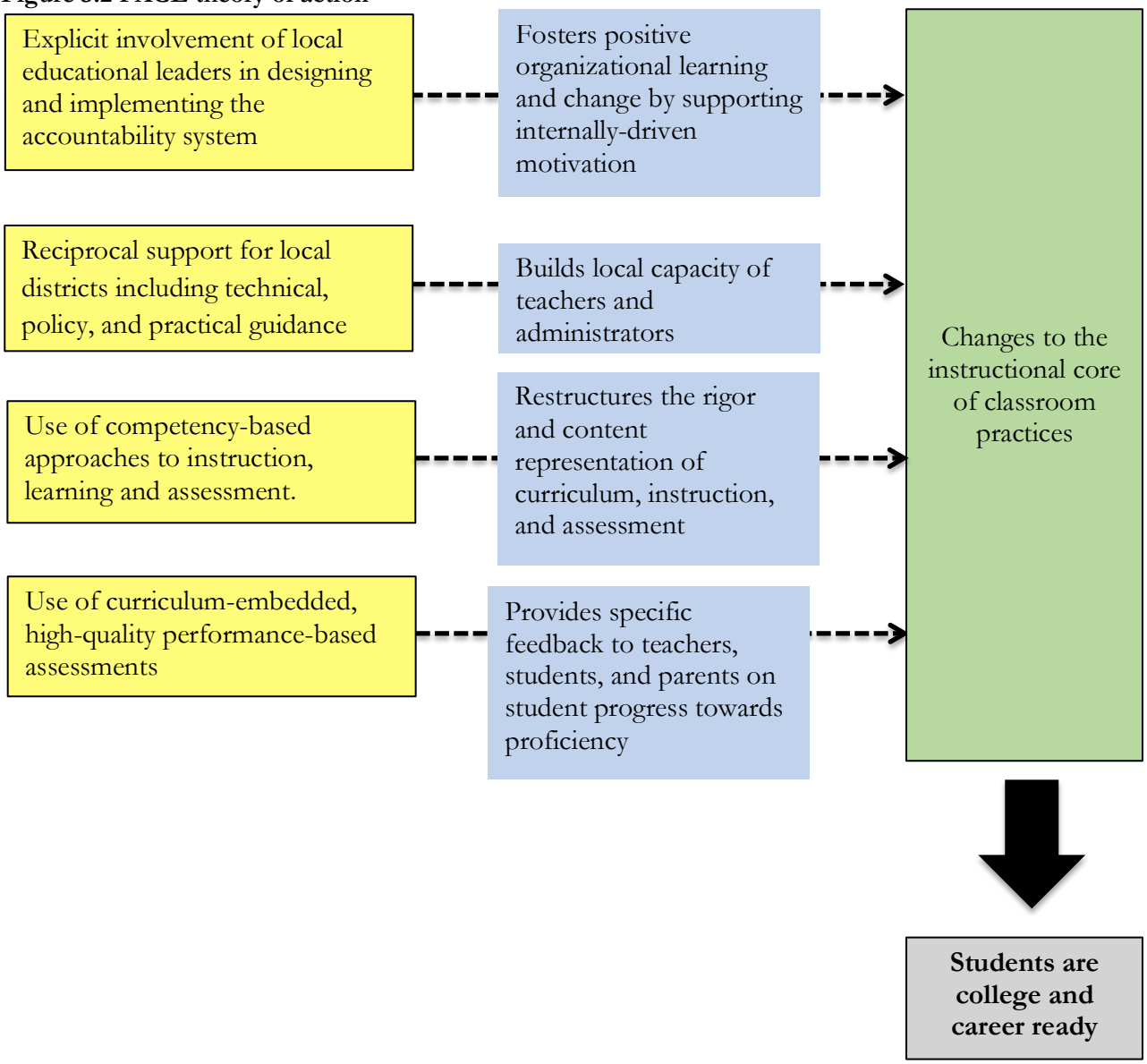
There are four main system design features with embedded assumptions of how those design features lead to changes in the instructional core of classroom practices. The first design feature is that local education leaders are explicitly involved in designing and implementing their own accountability system. This fosters positive organizational learning and change by supporting the internal motivation of educators. This contrasts with all-too-common top-down accountability and extrinsic approaches where the goals and methods of the accountability system are defined at

the state or federal levels and districts are simply expected to comply. The second design feature is that local education leaders are provided reciprocal support and capacity building to support their development of key capacities related to designing and implementing the system. This means the NH DOE and its technical partners provide high-quality professional development, training, and support to local districts in the technical, policy, and practical issues related to the system design and implementation. The third design feature is the use of competency-based approaches to learning, instruction, and assessment. These approaches structure learning opportunities for students to gain meaningful knowledge and skills at a depth of understanding that they can transfer to new real-world situations. These approaches also improve student motivation and engagement because they allow students more voice and choice in their own learning. The fourth design feature is the use of locally designed and curriculum-embedded performance assessments throughout the year. These high-quality assessments signal high learning expectations, monitor student learning, and provide specific feedback to teachers and students on their performance relative to the grade and subject competencies. Since these rich, cognitively demanding assessment experiences are curriculum-embedded, teachers can adjust their instruction in real-time to meet students where they are at and help them grow towards proficiency. The PACE Common Task serves as an exemplar for teachers of a high-quality performance assessment, rubric, and scoring protocols and procedures. As more PACE Common Tasks are designed, there is a bank of high-quality performance tasks and rubrics with anchor papers at different levels of performance to help drive positive instructional changes. The ultimate goal of PACE, as seen in the theory of action below, is that student achievement outcomes improve and that all students are college or career ready upon graduation from high school.

It is important to note, however, the State's *a priori* expectation for student achievement outcomes resulting from the PACE pilot over the first few years was "no harm" on Smarter

Balanced. This criterion was defined as "the performance of students in PACE districts does not decline compared to predicted scores on Smarter Balanced and actual trends on district interim assessments" (NHDOE, 2015, p. 5). The criterion of "no harm" on Smarter Balanced provides evidence that students in PACE schools/district were provided an equitable opportunity to learn the content standards because Smarter Balanced is aligned with the State's grade and subject level content standards.

Figure 3.2 PACE theory of action



A key premise of the PACE theory of action is that local education leaders are supported by the NHDOE and each other in creating the expertise necessary to implement the system with fidelity. There are many ways in which the PACE pilot builds local capacity both prior to and while implementing the PACE system. The following section provides a detailed description of the three-tiered system that prepares districts with the key capacities to implement the PACE system as intended. More information on local capacity building processes and collected validity evidence for the PACE system can be found in the PACE Technical Manual (Lyons, Evans, Marion, & Thompson, 2017).

Process for Districts to Implement PACE

The process for school districts[7] to be accepted for inclusion in the PACE pilot is based on a three-tiered system of rolling cohorts (NHDOE, 2015a). Districts are selected for participation in one of three rolling cohorts based on their application to the NHDOE, which includes a readiness survey related to competency-based education and performance-based assessment (NHDOE, 2016a). This process allows districts to enter at their current level of preparation and also helps the NHDOE identify areas of professional development support necessary for districts to become fully implementing PACE districts (M. Gfroerer, personal communication, November 21, 2016). This means districts do not have to enter at Tier 3; districts can skip Tiers 2 and 3 completely and just begin implementing as a Tier 1 district—it all depends on their level of readiness.

Table 3.4 provides specific definitions for each tier and an explanation of the targeted supports offered to districts by the NH DOE for each of the three tiers. Tier 1 districts are those districts that are implementing PACE. Tier 1 districts have reported implementing competency-based education in classrooms and have some experience and capacity with performance

[7] Although the term "school districts" or "districts" is used throughout this description, it is also the case that sometimes it is only one school within a district that has applied to join the tiered PACE structure or that a school administrative unit (SAU) comprised of multiple small school districts under one Superintendent has applied. For simplicity, I do not differentiate between these groups.

assessments of competencies. For those school districts that are not yet ready to move to Tier 1, the state provides targeted assistance to districts to help them move toward Tier 1 if they so choose. Tier 2 includes districts that report at least course level or school-wide competencies in place, but do not have a lot of experience with performance assessments. Tier 3 districts are at the "less advanced" development stage in terms of competency-based education and performance assessment and need more targeted assistance and support. In general, Tier 3 includes districts that report limited competency-based learning environments, do not implement competencies at the classroom level, and have no background with performance assessments.

Table 3.4 Definitions of PACE tiers with a description of the targeted support offered to districts by the NHDOE

	Definition	Targeted Support Provided by NHDOE to the District
Tier 1	Districts that are implementing the PACE pilot have reported implementation of local competencies in school-wide and classroom settings, and some experience with performance assessment in a competency-based learning environment. Evidenced a commitment to transitioning to implementing performance assessment of competencies for accountability purposes district-wide (K-12), and have articulated a beginning plan of how to best accomplish that transition in their community.	The district Superintendent and PACE team leader have the opportunity to meet monthly with PACE state-level leadership for policy and project management discussions. Access to workshop days throughout the year facilitated by experts, consultants, and coaches allowing cross-school learning of performance assessments within specific content areas and across grade-spans that support curriculum-embedded competency-based task design for formative and summative assessment purposes, scoring, and calibration. Coaching and guidance from experts in the development and implementation of common performance assessment tasks for accountability, based on readiness.
Tier 2	Districts that have reported to have course level and school-wide competencies in place and have at least some implementation of competencies in classroom settings. Competency-based learning environments may be evidence in some places in the district. Experience with task-based	Access to intense Quality Performance Assessment (QPA) training. Access to professional development from state and national experts on performance assessment literacy, beginning levels of performance task development, depth of knowledge levels, how to analyze at student work, reliable scoring, and

	performance assessment for competency attainment may be limited to extended learning opportunities or may not have been attempted in any systematic way.	local structures such as professional learning communities. Districts are also introduced to the NH PACE implementation protocols.
Tier 3	Districts that have reported no or few local active competency based learning environments, do not implement the competencies at the classroom level with students (though they may or may not have written competencies), and have no background experience with performance assessment of competencies.	Access to school-level coaching from state-contracted expert consultants on the topics of developing and implementing competencies and working with the state model competencies. Planning activities with other Tier 3 districts to prepare for greater involvement in performance assessment district-wide.

Note. Definitions and descriptions of targeted support taken from the PACE application (NHDOE, 2016a).

It is important to note that there is a continuum of district capacity related to competency-based education and performance assessment of competencies in each of the three tiers. For example, even within Tier 1, districts fall along a fidelity-of-implementation continuum (M. Gfroerer, personal communication, November 21, 2016). This developmental continuum may influence the direction and magnitude of the effects of the PACE pilot on student achievement outcomes investigated in this study. Although there is no formal investigation of this continuum yet available, the PACE state director provided an unofficial evaluation of Tier 1 district fidelity-of-implementation for the first two years of the PACE pilot (see Table 3.5)(M. Gfroerer, personal communication, November 21, 2016).

Table 3.5 Tier 1 district fidelity-of-implementation continuum for the first two years of the PACE pilot

Low Fidelity (Districts 165, 365, 705)	Mid-Level Fidelity (Districts 111, 439, 461)	High Fidelity (Districts 476 and 493)
District 165: Cohort 1 district; Rarely participated in PACE design decisions in 2013-2014 planning year; Struggles with consistency of implementation across schools because of varying degrees of administrator buy-in.	**District 111:** Cohort 2 district; Large district so it is unclear if implementation is consistent due to the time necessary for reform to trickle down to every classroom; Received Tier 2 and 3 level supports concurrently with Tier 1 implementation (not prior to joining Tier 1).	**District 476:** Cohort 1 district; Active participation in the PACE design process during the 2013-2014 planning year; Strong desire from district- and school-level administrators to implement PACE with fidelity, but teachers in their district may not have the same level of

		understanding or buy-in.
District 365: Cohort 2 district; Very small district; Joined Tier 1 without having done Tier 2 and 3 so their level of implementation has been effected by unfamiliarity.	**District 439:** Cohort 2 district; Rural district that has implemented competency-based education and performance assessments for years (e.g., was part of CBAS); Tier 2 district in Pilot Year 1	**District 493 (high school only):** Cohort 1 district; Active participation in PACE design process during the 2013-2014 school year; School was originally created as a competency-based school using performance assessments many years ago, which sometimes creates tension when teachers are asked to implement PACE in a certain way that may not be the same as the way they have been used to.
District 705: Cohort 2 district; Small K-8 charter school that is performance-based with an emphasis on the arts; Strong emphasis on project-based learning, but not competency-based; Joined Tier 1 without having gone through Tier 2 or 3.	**District 461:** Cohort 1 district; Active participation in the PACE design process during the 2013-2014 planning year; Large district so it is unclear if implementation is consistent due to the time necessary for reform to trickle down to every classroom.	

Note. Cohort 1 districts are those that began implementing PACE in Year 1 of the pilot (2014-2015 SY); whereas, Cohort 2 districts began implementing PACE in Year 2 of the pilot (2015-2016 SY).

In terms of Tier 2 and 3 districts, because districts can join the PACE tiers based on their point of preparation, some districts skip Tiers 2 and 3 and start in Tier 1. Table 3. 6 provides an overview of the districts in Tiers 2 and 3 during the first two years of the PACE pilot (2014-2016 school years).

Table 3.6 Description of Tier 2 and 3 districts in the first two years of the PACE pilot

District	Description	Tier Process
SAU 35	Collection of small districts under one Superintendent; K-12	Started in Tier 1 in 2016-2017 SY; did not start in Tier 2 or 3, although had completed QPA training on their own during the 2014-2016 school years, which is the targeted support offered to Tier 2 districts.
VLACS	Grades 6-12 online virtual charter school	Started in Tier 1 in 2016-2017 SY; is fully competency-based and performance-based; has completed QPA training in the past.
SAU 23	Collection of small districts under one Superintendent; K-12	Started in Tier 3 in 2015-2016 SY; moved to Tier 2 for 2016-2017 SY

SAU 58	Collection of small districts under one Superintendent; K-12	Started in Tier 3 in 2015-2016 SY; moved to Tier 2 for 2016-2017 SY
SAU 39	The elementary and middle school of one of the Year 1 Pilot implementing districts	Elementary school started in Tier 3 in 2015-2016 SY; Middle school started in Tier 2 in 2015-2016 SY; Both schools are in Tier 2 for 2016-2017 SY
SAU 60	Small K-12 rural school district	Started in Tier 2 in 2015-2016 SY and continue in Tier 2 in 2016-2017 SY
SAU 48	Elementary school only	Started in Tier 2 in 2015-2016 SY and continue in Tier 2 in 2016-2017 SY
SAU 37	One elementary school only (Parker Varney)	Started in Tier 2 in 2015-2016 SY and continue in Tier 2 in 2016-2017 SY
SAU 56	Small K-8 school district	Started in Tier 2 in 2015-2016 SY and continue in Tier 2 in 2016-2017 SY
SAU 43	All schools involved; K-12	Started in Tier 2 in 2016-2017 SY
SAU 53	K-8 school	Started in Tier 2 in 2016-2017 SY

Note. This table does not include Cohort 2 districts that became Tier 1 implementers in Year 2 of the pilot. Those districts are listed in Table 3.5; SAU=school administrative unit; SY=school year.

Population

The population includes all Grade 8 students in New Hampshire during the first two years of the PACE pilot (2014-2015 and 2015-2016 school years) who took a Smarter Balanced (SBAC) achievement test (N=26,936). The treatment group is Grade 8 students attending PACE schools who have been receiving competency-based instruction within a performance assessment accountability program for one or two years, depending upon which cohort their school is in (see Table 3.2). This means some Grade 8 students in the treatment group have been receiving treatment since Grade 6 (Cohort 1); whereas other Grade 8 students in the treatment group have been receiving treatment since Grade 7 (Cohort 2). All Grade 8 students in NH whether in the treatment or comparison group take the SBAC test at the end of Grade 8. Eighth grade was chosen because it is the only grade where PACE students have prior achievement test scores and take achievement tests in both ELA and math in the same year that are specifically aligned to the Common Core State Standards. NH competencies utilized by PACE implementing districts are aligned to the Common Core State Standards.

Datasets

State-level secondary datasets were merged to conduct these analyses. The first files contain student-level data on all students in New Hampshire (grades 3-11) who completed a spring 2015 or 2016 SBAC test in either math or ELA. The spring 2015 SBAC administration was the first operational administration of a Common Core State Standards aligned achievement test in New Hampshire. Variables in the files include identification variables such as research student IDs, school IDs, district IDs, grade level tested, and SBAC year id. SBAC vertical scale scores and achievement levels (1-4) for ELA and math were also provided. Other variables include student-level demographic information in a series of dichotomous variables indicating status related to: Individualized education plan (IEP)—a proxy for special education (yes/no); free and reduced lunch—a proxy for socio-economic status (yes/no); limited English proficient (LEP)(yes/no); and gender (male/female). A race variable provides information for seven different race/ethnic categories: American Indian or Alaskan Native; Asian; Black or African American; Hispanic or Latino; Native Hawaiian or Pacific Islander; Two or more races (non-Hispanic); or White. Almost 90% of the full, unweighted sample is White, consistent with the demographics of the state.

Another series of files contain student-level data on all students in New Hampshire (grades 3-11) who completed the New England Comprehensive Assessment Program (NECAP) achievement test in the Fall of 2012 or 2013 in either math or ELA. The Fall 2013 administration was the last administration of the NECAP assessment in New Hampshire. Similar variables are in the NECAP file as in the Smarter Balanced file.

Analytic Sample

There were five conditions for inclusion/exclusion in the analytic sample. First, Grade 8 students were removed from the sample if they attended schools with less than 10 students to eliminate possible data coding errors and students attending alternative schools (N=31; 0.1%). For example, there were some cases where a school had only one Grade 8 student because the school

was an out-of-state placement or alternative school. Second, Grade 8 students attending non-public schools (private or charter schools) or special education schools were removed prior to analyses (N=430; 1.6%). This allows for comparisons to be made between regular, public school students. Third, students who repeated Grade 8 were removed because effects may be systematically different for these students (N=18, 0.07%). Fourth, Grade 8 students attending any PACE Tier 2 or 3 schools (identified in Table 3.6) were removed so that an appropriate comparison group could be identified (N=2,617; 9.7%). As stated in the research questions, the comparison in this study is between PACE students and non-PACE students. Since PACE Tier 2 and 3 schools are receiving targeted supports from the state around competency-based education and performance assessments and there is a developmental continuum of implementation among all levels of schools in the PACE tier structure, it is important to remove students who attend schools receiving some level of treatment. Removing Tier 2 and 3 students from the sample ensures that the effect of the full treatment on student achievement outcomes can be investigated and compared to the effect of no treatment. There are too many confounds if PACE students' academic achievement scores are compared with other PACE tiered students that are attending schools receiving targeted support from the state, but have not yet chosen to implement PACE. And finally, Grade 8 students without prior achievement tests results and student background/demographic information were also removed from the sample by subject area (N=2,208 for math; N=2,225 for ELA; ~8%). Bias due to this type of listwise deletion is not likely because there is no reason to assume that this data is not missing completely at random. It is important to estimate achievement conditioned on prior achievement because past test performance is the most likely predictor of future test performance (Schmidt & Hunter, 1998). Students who fit the five inclusion/exclusion criteria were included in the unweighted analytic samples (N=21,632 for math; N=21,615 for ELA). There were 113 non-PACE schools and 7 PACE schools in each analytic sample.

Baseline Characteristics of the Unweighted Analytic Sample

Selection is at the district level because districts made the decision to implement the PACE pilot, not students or schools[8]. As a result, it is important to establish baseline equivalence for the PACE and non-PACE comparison groups in the analytic sample using district-level characteristics. In order to examine the district-level differences between the PACE and non-PACE comparison groups in the analytic sample, eight district-level characteristics were aggregated from NECAP data files by year (2012-13 or 2013-14) to capture pre-treatment differences in districts for those students in the analytic sample. These eight district-level characteristics are plausibly related to outcome and include: the percent of (1) male students in the district, (2) IEP students in the district, (3) free and reduced price lunch students in the district, (4) limited English proficient students in the district, (5) non-White students in the district, (6) students proficient or above in math on NECAP, (7) students proficient or above in ELA on NECAP, and (8) the number of students in the district. Since these variables were aggregated from the NECAP data files by year they only include students in grades 3-8 and 11. These district-level aggregated variables were then merged into the student-level analytic data file by district ID numbers and year so all Grade 8 students in one district have the same district-level percent by characteristic and year. An average[9] for each of the eight district-level characteristics was then computed by group/treatment status (PACE vs. non-PACE) using all the students in the analytic sample. Table 3.7 provides the baseline characteristics of the Grade 8 math (top panel) and Grade 8 ELA (bottom panel) analytic samples on district-level characteristics by treatment status (non-PACE vs. PACE). According to the *What Works Clearinghouse* Group Design Standards (Institute of Education Sciences, 2014), standardized mean differences in absolute value

[8] Note: In the first two years of the pilot, there are seven implementing districts with an 8th grade all of which have only one school with an 8th grade. This is true of most districts in NH because there are 120 schools with an 8th grade in the analytic sample, but only 113 districts.

[9] This is a weighted average because each district has a different number of students in the analytic sample each year.

between 0.00 and 0.05 "satisfies baseline equivalence", between 0.05 and 0.25 "requires statistical adjustment to satisfy baseline equivalence", and greater than 0.25 "does not satisfy baseline equivalence" between the treatment and comparison groups in the analytic sample (p. 15).

Table 3.7 Baseline characteristics of the unweighted Grade 8 math (top panel) and ELA (bottom panel) analytic samples on district-level characteristics by treatment status

		%male	%iep	%frl	%lep	%non-white	%math-prof	%ELA-prof	Nstud
Grade 8 Math									
Non-PACE	M	51.54	14.74	24.35	1.20	9.40	69.37	79.51	1454.55
N=20,018	*SD*	*2.08*	*3.17*	*14.38*	*1.60*	*7.53*	*9.99*	*7.26*	*1476.79*
PACE	M	51.31	16.45	34.91	2.37	10.59	62.34	72.83	1743.91
N=1,614	*SD*	*1.38*	*3.29*	*11.64*	*2.53*	*5.34*	*6.03*	*5.29*	*965.43*
	M Diff	0.23	-1.71	-10.56	-1.17	-1.19	7.04	6.68	-289.36
	SMD	0.13	-0.53	-0.81	-0.57	-0.19	0.88	1.06	-0.24
	p-value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Grade 8 ELA									
Non-PACE	M	51.55	14.74	24.35	1.19	9.40	69.37	79.51	1454.65
N=20,003	*SD*	2.07	3.17	14.38	1.60	7.53	9.99	7.26	1476.88
PACE	M	51.31	16.45	34.91	2.37	10.59	62.34	72.83	1743.31
N=1,612	*SD*	1.38	3.29	11.65	2.53	5.34	6.03	5.29	965.88
	M Diff	0.23	-1.71	-10.56	-1.17	-1.19	7.03	6.68	-288.66
	SMD	0.14	-0.53	-0.81	-0.57	-0.18	0.88	1.06	-0.24
	p-value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001

Note. All variables were aggregated from NECAP data files and merged into SBAC data file by district ID and year for each student in the analytic sample. %iep=percent of students with individualized education plans in the district; %frl=percent of students who qualify for free- and reduced-price lunch in the district; %lep=percent of students identified as limited English proficient in the district; %nonWhite=percent of students classified as American Indian/Alaskan Native, Asian, Black, Hispanic, Native Hawaiian/Pacific Islander, and Two or more races; %mathprof=percent of students proficient or above in math in the district; %ELAprof=percent of student proficient or above in ELA in the district; Nstud=total number of students in the district. M=mean; SD=standard deviation; M Diff=unstandardized difference in means; SMD=standardized mean difference using pooled standard deviations. SMD greater than 0.25 highlighted in yellow.

Overall, there are a few notable differences when comparing the PACE and non-PACE groups as evidenced by standardized mean differences greater than 0.25 (highlighted in yellow in the table above). First, there tends to be higher average district percentages of students receiving free- or reduced-priced lunch, IEP students, and limited English proficient students in the PACE group. There are also more students on average in PACE districts so the number of students in the PACE

group is higher than in the non-PACE group. Second, non-PACE districts tend to have higher percentages of students who are proficient or above in math and ELA than the PACE districts so the PACE group's average proficiency rate is lower.

There is no apparent reason why these pre-existing differences exist between the two groups. Districts are not financially incentivized to join PACE, but perhaps districts with higher levels of student need (broadly defined) are more likely to seek out assistance and capacity-building from the state to improve student achievement. Because five of the eight observed characteristics do not satisfy baseline equivalence (SMD>0.25), inverse propensity score weighting was employed to balance the two groups (PACE vs. non-PACE) on the observable district-level characteristics prior to outcome analyses.

Propensity Score Estimation

Propensity score methods allow a researcher to create equivalent treatment and comparison groups at baseline based on observable differences in the two groups so that unbiased estimates of average treatment effects can be made (Austin, 2011; Graham & Kurlaender, 2011; Guo & Fraser, 2015; Rosenbaum & Rubin, 1983). This method for identifying or weighting the analytic sample attempts to mimic a randomized control trial where participants receiving treatment are identical to the control group on observable characteristics so that the researcher can estimate unbiased treatment effects. This is important because without a randomized experimental design, which is often not possible in education contexts, selection bias can impact the estimates of treatment effects (Shadish, Cook, & Campbell, 2002). As already discussed, selection bias manifests itself in this study because school districts self-select into the PACE pilot. Unbiased estimates of average treatment effects are predicated on the assumption that there are no unobserved characteristics that predict assignment to treatment not included in the propensity score model (Rosenbaum & Rubin, 1985). One advantage of propensity score methods is that they reduce the dimensionality of the data into

one metric, which is useful when there are many observed differences that can bias treatment effects.

Propensity Score Models

A propensity score for student *i* is the conditional probability of being in the treatment group $(W_i = 1)$ versus the non-treatment group, given a vector of observed district-level covariates, x_i (Rosenbaum & Rubin, 1983): $e(x_i) = \text{pr}(W_i = 1|X_i = x_i)$. To estimate propensity scores, a binary logistic regression model was specified that included observed district-level characteristics plausibly related to outcome discussed above. These included percentage of students in the district who are designated as 1) qualifying for free- and reduced-price lunch, 2) receiving IEP services, 3) limited English proficient, 4) non-White, 5) proficient or above in math, and 6) proficient or above in ELA. Because district-level covariates were used in the logistic regression model for each student in the analytic sample, every student in a district by year had the same estimated propensity score. The distribution of propensity scores resulting from this model had good overlap and there were no propensity scores that fell outside the common support region (See Appendix B for additional information including parameter estimates from the propensity score model).

Guo and Fraser (2015) discuss different ways of using the propensity scores (or predicted probabilities) to reduce selection bias. These include nearest neighbor matching within a caliper and inverse propensity score weighting. Nearest neighbor matching uses the estimated probabilities from the logistic regression model and matches students from the PACE group to the non-PACE group if their estimated probabilities are within a certain caliper width. Inverse propensity score weighting uses a survey weighting approach to attempt to replicate a random experiment where each group (treatment and comparison) looks the same and their means are equal to the sample means (Guo & Fraser, 2015).

I decided to use inverse propensity score weighting (described in more detail below) because it does not result in trimming of the sample, which is especially important in this study because the trimming would have significantly reduced the number of schools/districts in the sample. The number of schools/districts in the sample would have been reduced because each student within a district by year has the same estimated probability of receiving treatment so nearest neighbor matching would have matched PACE students in one district to non-PACE students in only a limited number of districts with estimated probabilities that were most similar. For example, with 1:1 nearest neighbor matching with a caliper size set to $\varepsilon \leq 0.2\sigma_p$, where σ_p denotes standard deviation of the estimated propensity scores of the sample, the number of districts in the analytic sample goes from 113 to 38 per year. Since there is only one school with 8th grade in most districts in NH this would have reduced the number of schools from 120 to 40 per year. Reducing the number of districts/schools in the analytic sample reduces the variation in outcome that can be explored at multiple levels, which then makes the parameter estimates less precise.[10]

Inverse propensity score weighting. There are two different inverse propensity score weights that can be calculated: average effect of treatment on the treated (ATT) and average treatment effect (ATE). I calculated the ATE weight since my research question focuses on comparing the PACE group to the non-PACE group. The equation for creating the ATE weight is as follows:

Treated = $\frac{1}{\hat{e}(x_i)}$ Comparison = $\frac{1}{1-\hat{e}(x_j)}$

where $\hat{e}(x)$ is the estimated propensity score for each treated student *i* or comparison student *j*.

[10] I did examine the standardized mean differences between the PACE and non-PACE comparison group on the baseline characteristics using 1:1 nearest neighbor matching with a 0.25 caliper and found that this procedure did not provide better balance between the groups on the observed district-level characteristics. Increasing the number of matches possible did not result in significantly less trimming at the district/school-level.

PACE students with large estimated probabilities of being assigned to treatment due to their district-level characteristics have small ATE weights (e.g., 1.3), whereas PACE students with small estimated probabilities of being assigned to treatment due to their district-level characteristics have large ATE weights (e.g., 34). On the other hand, non-PACE students with large estimated probabilities of being assigned to treatment due to their district-level characteristics have medium ATE weights (e.g., 5.6), whereas non-PACE students with small estimated probabilities have small ATE weights (e.g., 1). This survey weighting approach, as mentioned earlier, attempts to balance the two groups (PACE vs. non-PACE) on observable characteristics related to both selection and outcome so that unbiased estimates of treatment effects can be made. It does so by weighting down treatment students with large estimated probabilities and comparison students with small estimated probabilities and vice versa (weighting up treatment students with small estimated probabilities and comparison students with large estimated probabilities).

Guo and Fraser (2015) recommend using weighted least squares regression (with continuous covariates) and weighted logistic regression (with dichotomous variables) to examine covariate balance. I also used standardized mean differences to examine whether the inverse propensity score weighting satisfied baseline equivalence (Austin, 2011). The ATE weight was then used as a probability weight in subsequent statistical analyses.

One disadvantage of this approach is that some individual-level weights may be really large. However, there is also a corrected version of the ATE weight that can be used if some of the individual-level weights are really large. The corrected version of the ATE weight basically multiples the inverse propensity score by a constant—a process called stabilization.

ATE Weight (Corrected Version)

$$\text{Treated} = \frac{\sum_{i=1}^{n_1} \hat{e}(x_i)}{n_1} * \frac{1}{\hat{e}(x_i)} \qquad \text{Comparison} = \frac{\sum_{j=1}^{n_0} [1-\hat{e}(x_j)]}{n_0} * \frac{1}{\hat{e}(x_j)}$$

I found that the corrected version of the ATE weight reduced the range and standard deviation of the ATE weight as desired, but did not produce better balance on the district-level characteristics between the two groups. Also, the corrected version of the ATE weight when used in subsequent analyses inflated the intraclass correlation coefficient considerably. For these reasons, the corrected version of the ATE weight was not used.

Baseline Characteristics of the Inverse Propensity Score Weighted Analytic Sample

Table 3.8 provides baseline characteristics of the Grade 8 inverse propensity score weighted math and ELA analytic samples on observed district-level characteristics by treatment status. The absolute value of the standardized mean differences for each district-level variable should fall below 0.25 if equivalence between the two groups has been established at baseline (Institute of Education Sciences, 2014). Even after inverse propensity score weighting, there are differences between the two groups on district-level characteristics. In almost all cases, the PACE group differs in ways that might underestimate the effect of treatment rather than overestimate. For example, the PACE group comes from districts with higher percentages of IEP students and students who qualify for free- and reduced-price lunch. The non-PACE group comes from districts that have higher percentages of students who are proficient or above in math or ELA. And yet in comparison to the unweighted Grade 8 math and ELA analytic sample, there is only one standardized mean difference above 0.25. This implies that the inverse propensity score weight is creating more equivalent groups.

Table 3.8 Baseline characteristics of the inverse propensity score weighted Grade 8 math (top panel) and ELA (bottom panel) analytic samples on district-level characteristics by treatment status

		%male	%iep	%frl	%lep	%non-white	%math-prof	%ELA-prof	Nstud
Grade 8 Math									
Non-PACE	M	51.47	14.92	25.63	1.45	9.83	68.49	78.67	1551.55
N=22,078	*SD*	2.56	3.29	15.04	2.16	8.21	10.59	7.96	1680.09
PACE	M	51.14	15.73	28.41	1.19	8.60	66.26	76.67	1253.05
N=16,147	*SD*	1.84	3.55	12.87	1.69	4.54	8.93	6.90	885.98
	M Diff	0.33	-0.81	-2.79	0.26	1.22	2.23	2.00	298.50
	SMD	0.15	-0.24	-0.20	0.14	0.19	0.23	0.27	0.23
	p-value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Grade 8 ELA									
Non-PACE	M	51.47	14.92	25.63	1.45	9.83	68.50	78.67	1551.81
N=22,063	*SD*	2.56	3.29	15.03	2.16	8.21	10.59	7.96	1680.59
PACE	M	51.14	15.74	28.43	1.18	8.61	66.24	76.66	1253.78
N=16,147	*SD*	1.84	3.55	12.87	1.68	4.54	8.92	6.90	886.04
	M Diff	0.34	-0.82	-2.80	0.27	1.22	2.25	2.01	298.03
	SMD	0.15	-0.24	-0.20	0.14	0.19	0.23	0.27	0.23
	p-value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001

Note. ATE weights applied. All variables were aggregated from NECAP data files and merged into SBAC data file by district ID and year. %iep=percent of students with individualized education plans in the district; %frl=percent of students who qualify for free- and reduced-price lunch in the district; %lep=percent of students identified as limited English proficient in the district; %nonWhite=percent of students classified as American Indian/Alaskan Native, Asian, Black, Hispanic, Native Hawaiian/Pacific Islander, and Two or more races; %mathprof=percent of students proficient or above in math in the district; %ELAprof=percent of student proficient or above in ELA in the district; Nstud=total number of students in the district. M=mean; SD=standard deviation; M Diff=unstandardized difference in means; SMD=standardized mean difference using pooled standard deviations. SMD greater than 0.25 highlighted in yellow.

These differences between the PACE treatment and non-PACE comparison group even after inverse propensity score weighting suggest that any findings resulting from subsequent multivariate analyses should be considered descriptive rather than causal. Students came from districts that differed in these observed ways, but also likely in unobserved ways that were related to both their treatment status and measured student achievement outcomes. For example, only a limited number of covariates was included in the propensity score model.

Student-level characteristics prior to outcome analyses were fairly equivalent between the PACE and non-PACE comparison group (Table 3.9), although there was about 31% of a standard

deviation more limited English proficient students in the PACE group. Similar to the district-level characteristics, in almost all cases, the PACE group differs in ways that might underestimate the effect of treatment rather than overestimate. For example, the PACE group tends to have students with slightly lower prior achievement, more limited English proficient students, and a lower percentage of students proficient or above in Grade 6 math and ELA.[11]

Table 3.9 Student-level baseline characteristics of the weighted Grade 8 math (top panel) and ELA (bottom panel) analytic samples by treatment status

		Gr 6 necap	%male	%iep	%frl	%lep	%non White	%Gr 6 math Prof	%Gr 6 ELA prof
Grade 8 Math									
Non-PACE	M	646.24	51.45	13.68	24.05	0.64	9.10	74.15	80.37
N=22,078	*SD*	*11.92*	*0.50*	*0.34*	*0.43*	*0.08*	*0.29*	*0.44*	*0.40*
PACE	M	645.36	49.80	14.32	38.27	0.56	5.32	71.47	77.29
N=16,147	*SD*	*11.40*	*0.50*	*.35*	*0.49*	*0.08*	*0.22*	*0.45*	*0.42*
	M Diff	0.88	0.01	-0.01	-0.14	0.00	0.04	0.03	0.03
	SMD	0.08	0.02	-0.02	-0.31	0.01	0.15	0.06	0.08
	p-value	<.001	<.05	>.05	<.001	>.05	<.001	<.001	<.001
Grade 8 ELA									
Non-PACE	M	649.35	51.45	13.67	24.03	0.62	9.07	74.17	80.37
N=22,063	*SD*	*12.07*	*0.50*	*0.34*	*0.43*	*0.08*	*0.29*	*0.44*	*0.40*
PACE	M	648.63	49.78	14.35	38.26	0.49	5.25	71.50	77.26
N=16,147	*SD*	*12.37*	*0.50*	*0.35*	*0.49*	*0.07*	*0.22*	*0.45*	*0.42*
	M Diff	0.73	0.02	-0.01	-0.14	0.00	0.02	0.03	0.02
	SMD	0.06	0.03	-0.02	-0.31	0.02	0.15	0.06	0.08
	p-value	<.001	<.05	>.05	<.001	>.05	<.001	<.001	<.001

Note. Student-level demographic variables are from SBAC data files. necap=mean Grade 6 prior achievement on NECAP test for math or ELA depending upon the sample; %iep=percent of students with identified disabilities; %frl=percent of students who qualify for free- or reduced-price lunch; %lep=percent of students who are limited English proficient; %nonwhite=percent of students designated as American Indian, Black, Asian, Hispanic, Pacific Islander, or two or more races; %mathprof=percent of students proficient or above on Grade 6 math NECAP; %ELAprof=percent of students proficient or above on Grade 6 ELA NECAP. M=mean; SD=standard deviation; M Diff=unstandardized difference in means; SMD=standardized mean difference using pooled standard deviations. SMD greater than 0.25 highlighted in yellow.

[11] District- and school-level baseline characteristics were also examined using three groups non-PACE, PACE (1 yr) and PACE (2 yrs) because of the multivariate modeling approach employed next. The standardized mean differences are very similar to what is reported in this section and therefore are not repeated. The standard practice is to report out on the binary logistic regression model.

Measures

Four different types of variables are included in the analyses: (1) identification (ID) variables, (2) student-level outcome variables, (3) student-level control variables, and (4) school-level treatment and control variables. Because some districts in NH and some districts in the PACE pilot only have one school in the district (e.g., Districts 365, 705, 493), only student- and school-level measures were included in the outcome analyses.

Level 1: Student-Level Outcome Variables

ELA and math achievement. The Smarter Balanced Assessment Consortium (SBAC) achievement tests are used to measure 8th grade student achievement in ELA and math (subject areas were modeled separately). Scale scores fall on a continuous scale from approximately 2000 to 3000. Students fall into one of four achievement levels based on their scale scores. Table 3.10 below shows the range of scale scores for each achievement level for math and ELA. Students performing at Levels 1 and 2 are considered below proficient. Levels 3 and 4 are considered proficient or above—in other words, "on track to demonstrating the knowledge and skills necessary for college and career readiness" (Smarter Balanced Assessment Consortium, n.d.).

Table 3.10 Range of Grade 8 Smarter Balanced scale scores for each achievement level by subject area

	Level 4	Level 3	Level 2	Level 1
Mathematics	>2652	2586-2652	2504-2585	<2504
ELA/Literacy	>2667	2567-2667	2487-2566	<2487

Note. Scale score ranges for each achievement level taken from (Smarter Balanced Assessment Consortium, n.d.).

The SBAC technical report (2015) provides detailed tables explicating the essential validity evidence gathered for the computer adaptive, summative assessments and how those pieces of evidence support the argument that the assessments are indeed measuring what they purport to measure; namely, student achievement in ELA/literacy and mathematics relative to the Common

Core State Standards. SBAC is an assessment used to determine if and how well students are progressing towards college and career readiness. One advantage of the SBAC tests is that the test blueprint lends itself to measuring higher-order thinking and problem solving skills (Herman & Linn, 2013). This means that SBAC measures the breadth and depth of the content standards and is therefore a fair outcome measure for both the treatment and comparison group.

Level 1: Student-Level Control Variables

Prior ELA and math achievement. Each Grade 8 students' prior ELA or math achievement is their individual-level Grade 6 NECAP scale score from either Fall 2012 or Fall 2013 (see Table 3.11). The NECAP achievement test was aligned with the NH Content Frameworks that were in place before the Common Core State Standards were adopted. It is important to note that Grade 6 NECAP scores are intended to measure student achievement on 5th grade performance standards since the NECAPs were administered in the fall of each year. The reason why SBAC prior achievement scores cannot be used is because SBAC was administered in NH for the first time in Year 1 of the pilot (2014-2015), which means the only prior achievement scores available for students in Year 1 and for Cohort 1 students in Year 2 are from the NECAP test.[12] Modeling the two years of the pilot together allows for more robust comparison across years and the estimation of dosage effects based on the number of treatment years. Predicting student achievement on Grade 8 SBAC using Grade 6 NECAP as a predictor variable does not assume that the two score scales are comparable. Instead, I demonstrate that there is a linear relationship between student performance on the NECAP assessment that can be used to predict variance in student performance on the SBAC assessment. Preliminary analyses support this assumption since there is a very strong linear relationship between the two assessments for Grade 8 students ($r=.79$, $p<.001$). All prior

[12] The use of a different achievement test also precludes certain analytic methods and designs such as interrupted time series because the outcome achievement test is not designed to measure the same content standards.

achievement variables were grand-mean centered prior to analysis in order to aide interpretation of the intercept (Raudenbush & Bryk, 2002)[13].

Student-level demographic variables. Analyses controlled for student-level factors that may affect student achievement using five dummy variables: free/reduced lunch (frl=1 or 0), individualized education plan status (iep =1 or 0), and gender (male=1 or 0). Limited English proficiency status (lep =1 or 0) and race/ethnicity (non-White =1; White=0)[14] were also examined but ultimately not included because of the low percentage of students in the analytic sample with those characteristics, which mirrors the state demographics.

Level 2: School-Level Treatment and Control Variables

Schools are also nested within school districts in this study; however only a 2-level model was specified because there is typically one school per district with 8th graders so school-level effects and district-level effects are confounded. I chose to model school-level treatment and controls instead of district-level treatment and controls because conceptually it is more likely that variation in individual student achievement is more affected by peer effects within school rather than peer effects within district. For example, it seems plausible that some variation in Grade 8 student-level achievement is explained by the performance of Grade 6 and Grade 7 students attending the same school more than student achievement in the entire district.

School mean prior ELA and math achievement. One school-level predictor of student achievement may include how their peers performed in similar subjects in the preceding year, otherwise known as peer effects (Hanushek, Kain, Markman, & Rivkin, 2001). For example, do students attend a school where their peers tend to perform really well on ELA and/or math achievement tests, or do they attend a school where their peers tend to perform really poorly on

[13] It is more common to group-mean center level-1 variables, however, the between-school variance went up from the unconditional/null model when this variable was group-mean centered.
[14] A dichotomous variable contrasting non-White vs. White students is included for race/ethnicity because the sample is almost 90% White.

ELA and/or math achievement tests? This school-level control variable (pctmathprof or pctELAprof) aggregates the percent of students who were proficient or above on the NECAP in ELA or math separately by school in order to create a school mean prior achievement measure. The computed variables were grand-mean centered to aide interpretation of the intercept.

Additional school-level control variables. SBAC data files were used to compute these measures. Analyses controlled for percent of students in the school who receive free/reduced lunch (pctfrl) and the number of students in the school (Nkids). These school-level control variables were grand-mean centered $(\text{e.g.}, PCTFRL_j - \overline{PCTFRL_J})$ to aide interpretation of the intercept. A dummy variable for SBAC year (sbacid) was also included since Grade 8 students in 2014-15 and 2015-16 are analyzed together. I also examined the percentage of students in the school who have an individualized education plan (pctiep), designated as limited English proficient (pctlep), and non-White (pctnonwhite) as other potential control variables, but ultimately removed them because of either low mean percentages of students within schools or poor model fit.

Treatment variables. In order to examine whether there are non-linear treatment and/or dosage effects, treatment effects were modeled using two dummy variables (Table 3.11). The first dummy variable (treat1) indicates whether a PACE school was in its first year of implementation or not in the 2014-15 or 2015-16 school year. The second dummy variable (treat2) indicates whether a PACE school was in its second year of implementation or not in the 2015-16 school year. All non-PACE comparison schools were coded as "0".

Table 3.11 Outcome, prior achievement, and treatment status variables by pilot year

	Treatment Group		Comparison Group
Pilot Year	**Cohort 1**	**Cohort 2**	**All**
2014-15 School Year			
8th grade outcome:	treat1	0	0
SBAC Spring 2015 ELA and Math	(n_students=477)	(n_students=381)	(n=9,954)
6th grade prior achievement:	(n_schools=3)	(n_schools=4)	(n_schools=113)
NECAP Fall 2012 ELA and Math			
2015-16 School Year			
8th grade outcome:	treat2	treat1	0
SBAC Spring 2016 ELA and Math	(n_students=456)	(n_students=300)	(n_students=10,064)
6th grade prior achievement:	(n_schools=3)	(n_schools=4)	(n_schools=113)
NECAP Fall 2013 ELA and Math			

As you can see from Table 3.11, this means that there were three PACE schools with 456 students who received two years of treatment and seven PACE schools with 681 students who received one year of treatment.

Analytic Approach

To address the research question about the extent to which PACE students differ from non-PACE students with similar probabilities of being selected into treatment in terms of their student achievement outcomes, the first part of the analyses focused on descriptive and exploratory analyses. Mathematics and ELA were explored separately. Bivariate plots and OLS regression analysis were used to explore and estimate the relationship between the level-1 predictors and SBAC student achievement for PACE and non-PACE students. In addition, the relationship between estimated intercepts and slopes from level-1 OLS regression and the level-2 predictors was explored.

Then, in order to answer research question #1, analyses focused on estimating the average treatment effect of the PACE pilot. Since students are nested within schools, multi-level modeling (Raudenbush & Bryk, 2002) was used to estimate the average treatment effect of the NH PACE pilot on the student-level outcome variables. ELA and math achievement were modeled separately and student- and school-level predictors and controls were used to account for differences in

students' demographic/ background characteristics and baseline test scores. Multi-level modeling handles the school-based clustering of achievement by distinguishing between-school variation from within-school variation. It thus allows the estimation of the effects of level-1 predictors to vary over level-2 predictors, and for the testing of cross-level effects on the outcome variables.[15]

In order to address research question #2, I examined whether treatment effects vary according to student-level characteristics such as prior achievement, free- and reduced-price lunch status, disability status, and gender. This provided insight into whether certain subgroups of PACE students are differentially affected by treatment. It is important to test for effects by subgroups because achievement gaps may be exacerbated, reduced, or remain the same for certain subgroups and not others. Due to the low percentage of minority and limited English proficient students in the Grade 8 NH population, this study cannot provide any insight into achievement gaps by race/ethnicity or limited English proficient status.

In order to address research question #3, I examined how treatment effects differ between PACE schools using the level-2 random effect estimates. I compared the predicted school-level achievement outcomes from the preferred multi-level model specification to observed differences in mean SBAC school-level performance. I was interested in the extent to which PACE schools performed better than predicted (positive residuals) or worse than predicted (negative residuals), as well as if there were any patterns across PACE schools or pilot years.

One advantage of this analytic approach is that insofar as the propensity score estimation model or the multi-level regression model is correctly specified, the combination of the two provides a doubly robust estimate of average treatment effects (Funk et al., 2011). This analytic approach attempts to account for pre-existing differences between the treatment and comparison groups in

[15] Schools are also nested within school districts in this study; however only a 2-level model was used because there is one school per district so school-level effects and district-level effects are confounded.

two ways. First, it adjusts for how observed district-level differences are associated with selection using propensity score weighting. Second, it adjusts for how observed student and school differences are associated with outcome using multi-level regression models. If either of the methods accurately accounts for pre-existing student-, school-, or district-level differences, the treatment effect estimates may be unbiased. However, as mentioned before, since the two groups were not equivalent at baseline on the district-level characteristics included in the propensity score model, all results should be interpreted as observational, not causal.

The multi-level analyses followed five steps. First, I fit an unconditional model separately for each outcome variable by subject area. This allowed me to estimate the intraclass correlation of the outcome to gauge the amount of variation in ELA and math achievement that occurs within-schools as opposed to between-schools. A fully unconditional two-level model that predicts student achievement in Grade 8 math was specified as follows:

Model 0: Fully Unconditional Model

Level 1: Student Level (Within school analysis)
$SBAC\_\text{Math}_{ij} = \beta_0 + r_{ij}$
where $r_{ij} \sim N(0, \sigma^2)$

Level 2: School Level (Between school analysis)
$\beta_0 = \gamma_{00} + u_{0j}$
where $u_{ij} \sim N(0, \tau_{00})$

Unconditional Composite Model
$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$

Estimated fixed effects:

1. Intercept, $\hat{\gamma}_{00}$=estimated average Grade 8 school math achievement.

Estimated random effects:

1. Level-1 variance, $\hat{\sigma}^2$=population variance of Y_{ij} among students within schools, or the estimated within school variance.
2. Level-2 intercept variance, $\hat{\tau}_{00}$=population variance in intercepts across schools, or the estimated between school variance.

Estimated intraclass correlation coefficient (ICC) equation:

$$\hat{\rho} = \frac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \hat{\sigma}^2} = ICC \quad (1)$$

The intraclass correlation coefficient is the percent of total variation in student achievement that occurs between schools; the rest occurs within schools. The level-2 intercept variance (τ_{00}) effectively places an upward bound or ceiling on the amount of variation in student achievement that can ever be "explained" by school-level (level-2) predictors. Including level-2 predictors into the model hopefully reduces the size of this between-school variance component so that I have "explained" part of the explainable variation between schools with regards to student achievement. Similarly, including additional level-1 predictors into the model hopefully reduces the size of the within-school variance component so that I have "explained" part of the explainable variation within schools with regards to student achievement.

Second, I fit a series of models with level-1 control variables only. I started with a model that includes all level-1 covariates, but no level-2 covariates. I added level-1 covariates one at a time, testing each as fixed effects only. Random effects for the level-1 control variables were not included in the model because the models did not converge either in this step or in later steps. This is most likely due to the dichotomous nature of most of the level-1 control variables. The goal was to specify the most parsimonious measurement model at level-1 to combine with the most parsimonious measurement model at level-2. Below I present an equation of a model that predicts Grade 8 math student achievement with all the level-1 covariates.

Model 1: Model with all level-1 control variables included.

Level 1: Student Level (Within School Analysis)

$$SBAC\_Math_{ij} = \beta_{0j} + \beta_{1j}(NECAP_{ij}) + \beta_{2j}(FRL_{ij}) + \beta_{3j}(IEP_{ij}) + \beta_{4j}(MALE_{ij}) + r_{ij}$$

where $r_{ij} \sim N(0, \sigma^2)$

Level 2: School Level (Between School Analysis)

$\beta_{0j} = \gamma_{00} + u_{0j}$

$\beta_{1j} = \gamma_{10}$

$\beta_{2j} = \gamma_{20}$

$\beta_{3j} = \gamma_{30}$

$\beta_{4j} = \gamma_{40}$

where $u_{0j} \sim N(0, \tau_{00})$

Composite Model:

$Y_{ij} = [\gamma_{00} + \gamma_{10}NECAP_{ij} + \gamma_{20}FRL_{ij} + \gamma_{30}IEP_{ij} + \gamma_{40}MALE_{ij}] + [u_{0j} + r_{ij}]$

Estimated fixed Effects:

1. Intercept, $\hat{\gamma}_{00}$
2. NECAP, $\hat{\gamma}_{10}$
3. FRL, $\hat{\gamma}_{20}$
4. IEP, $\hat{\gamma}_{40}$
5. MALE, $\hat{\gamma}_{50}$

Random Effects:

There are 2 random effects in this model: an estimated level-1 within-school residual variance and an estimated level-2 between-school residual variance.

A key question that I explored with this model is how much of the within-school variance in Grade 8 math achievement is "explained" by one or more of the level-1 control variables. To do so, I compared estimates of σ^2 from the unconditional (*u*) and conditional models (*c*) using the following equation:

$$\frac{\sigma^2_u - \sigma^2_c}{\sigma^2_u} \quad (2)$$

In this way, I explored how much within-school variation in student Grade 8 math achievement was "explained" by adding different level-1 predictors to the model—a pseudo-R^2 statistic.

Third, I fit a series of models with level-2 treatment and control variables only. The goal was to find the most parsimonious "means as outcomes" model prior to fitting models with both level-1 and level-2 variables. The mean is included as an outcome in this model because the $\hat{\beta}_{0j}$ estimates are school mean math achievement using school-level predictors.

Model 2: "Means as Outcomes" Model with all Level-2 predictors

Level-1: Within School Analysis

$$SBAC\_Math_{\mathrm{ij}} = \beta_{0\mathrm{j}} + r_{ij}$$

where $r_{ij} \sim N(0, \sigma^2)$

Level-2: Between School Analysis

$$\beta_{0\mathrm{j}} = \gamma_{00} + \gamma_{01}(SBACID_j) + \gamma_{02}(PCTMATHPROF_j) + \gamma_{03}(PCTFRL_j) + \gamma_{04}(PCTIEP_j) + \gamma_{05}(Nkids_j) + \gamma_{06}(TREAT1_j) + \gamma_{07}(TREAT2_j) + \gamma_{08}(TREAT1_j * SBACID_j) + u_{0j}$$

where $u_{0j} \sim N(0, \tau_{00})$

Composite model:

$$Y_{\mathrm{ij}} = \gamma_{00} + \gamma_{01}(SBACID_j) + \gamma_{02}(PCTMATHPROF_j) + \gamma_{03}(PCTFRL_j) + \gamma_{04}(PCTIEP_j) + \gamma_{05}(Nkids_j) + \gamma_{06}(TREAT1_j) + \gamma_{07}(TREAT2_j) + \gamma_{08}(TREAT1_j * SBACID_j) + u_{0j} + r_{\mathrm{ij}}$$

Estimated fixed effects:

1. Intercept, $\hat{\gamma}_{00}$=school mean Grade 8 math achievement when all other level-2 covariates are set to zero.

2. SBACID, $\hat{\gamma}_{01}$=the effect of SBAC year ID on Grade 8 school mean math achievement.

3. PCTMATHPROF, $\hat{\gamma}_{02}$=the effect of school percent math proficient or above on Grade 8 school mean math achievement.

4. PCTFRL, $\hat{\gamma}_{03}$=the effect of percentage of students who qualify for free- and reduced-price lunch in the school on Grade 8 school mean math achievement.

5. PCTIEP, $\hat{\gamma}_{04}$=the effect of percentage of students who have an IEP plan in the school on Grade 8 school mean math achievement.

6. Nkids, $\hat{\gamma}_{05}$=the effect of the number of students in the school on Grade 8 school mean math achievement.

7. TREAT1, $\hat{\gamma}_{06}$=the effect of one year of treatment on Grade 8 school mean math achievement.

8. TREAT2, $\hat{\gamma}_{07}$=the effect of two years of treatment on Grade 8 school mean math achievement.

9. TREAT1*SBACID, $\hat{\gamma}_{08}$=the effect of one year of treatment varies by year of SBAC test on Grade 8 school mean math achievement.

Estimated random effects:

1. Level-1 variance, $\hat{\sigma}^2$
2. Level-2 intercept variance, $\hat{\tau}_{00}$

Because this model has no level-1 predictors in the model, I explored how much parameter variance I explained by adding the level-2 variables. This was done in two ways. First, I subtracted the fully unconditional model variance component for $\hat{\tau}_{00}$ from this model's variance component for $\hat{\tau}_{00}$ and then divided by the unconditional model's $\hat{\tau}_{00}$. This answered the question: how much of the "explainable" parameter variance in Grade 8 school mean math achievement have I explained by adding the level-2 variables in the model? The resulting estimate acts as a pseudo-R^2 statistic.

A second way I explored how much more parameter variance was left to explain was to calculate a conditional ICC. A conditional ICC, or residual ICC, is the intraclass correlation among comparable schools. This ICC, as before, is the portion of total variance that occurs at the school level, but now this estimate is conditional on the level-2 variables being in the model. The conditional ICC is specified using the same formula as the unconditional ICC.

Fourth, I fit a multilevel model specification with both level-1 and level-2 predictors for each outcome variable using the most parsimonious models from Models 1 and 2 (Model 3). I then added the cross-level effects between treatment status and level-1 predictors (Model 4). I have not specified any of those models below because it depends upon the results of the analyses above. Lastly, I conducted residual analysis to evaluate the tenability of the "final" model's assumptions and then interpreted and explained all parameter estimates.

Summary of Analytic Approach

The main predictors of interest for research question #1 are the set of two treatment dummy variables, which indicate the effects associated with the number of treatment years, and the interaction between treatment and year. The first, treat1, is set to 1 if a school is in its first year of PACE implementation in either 2014-15 or 2015-16. The second, treat2, is set to 1 if a school is in its second year of PACE implementation in 2015-16. The interaction between treat1 and SBAC year ID is included to examine whether treatment effects differ by treatment year since treat1 includes both the 2014-15 and 2015-16 school years. There is no interaction between treat2 and SBAC year ID because schools implementing PACE for two years only took the SBAC test in the 2015-16 school year. The parameter estimates associated with these treatment dummy variables and interaction answer the first research question about average treatment effects in math and ELA across the first two years of the pilot.

In order to answer research question #2, the cross-level interactions between the treatment dummy variables (treat1 and treat2) and the level-1 student characteristics (necap, frl, iep, and male) were examined. These parameter estimates provide insight into whether certain subgroups of students differentially benefit or are "harmed" by participating in the PACE pilot.

In order to answer the third research question, the level-2 intercept residuals associated with the "final" model were computed and analyzed for PACE schools by treatment year. Specifically, I examined whether PACE schools outperform or underperform their predicted mean school achievement. These analyses provide insight into the extent to which treatment effects vary among PACE schools and school years, as well as if there are any patterns in performance based upon the informal conversation about each district's fidelity-of-implementation communicated by the NHDOE (see Table 3.5).

Summary

In this chapter, I provided a detailed overview of the study context and treatment. I also described the datasets, population, and the propensity score methods used to identify the analytic sample. I then explained the outcome measures alongside the level-1 and level-2 predictors and control variables used. The analytic approach was detailed step-by-step so that another researcher could replicate this study. The chapter ended with a brief overview of how I would use the analytic output to answer the research questions. Chapter Four presents a detailed overview of findings from this analytic approach.

Chapter 4: Findings

In this chapter, I report the study's findings. The chapter is organized by subject area—math first followed by ELA—and according to the three research questions. To address the first research question, I examined the average treatment effect of the PACE pilot on Grade 8 student achievement outcomes in comparison to non-PACE students with similar probabilities of being selected into treatment. To address the second research question, I investigated whether average treatment effects vary according to student-level characteristics such as prior achievement, free- and reduced-price lunch status, IEP status, and gender. To address the third research question, I examined variation in treatment effects among PACE schools comparing observed vs. predicted mean SBAC Grade 8 achievement at the school-level. The chapter concludes with a summary of findings synthesized across subject areas.

Math

Descriptive Analyses

To address the research question about the extent to which PACE students differ from non-PACE students with similar probabilities of being selected into treatment in terms of their student achievement outcomes in math, the first part of the analyses focused on descriptive and exploratory analyses. Descriptive statistics and distributions of all the variables were examined (see Table 4.1). There were small percentages of students classified as limited English proficient (LEP 1%) and non-White (8%) in the sample, which is consistent with the state's demographics.

Table 4.1 Descriptive statistics on variables in the inverse propensity score weighted Grade 8 math sample (wtd. N_students=38,225)

	Minimum	Maximum	Mean	SD
necap	600	680	645.87	11.712
gend	0	1	.51	.500
iep	0	1	.14	.346
frl	0	1	.30	.459
lep	0	1	.01	.078
nonwhite	0	1	.08	.26341
pctmathprof	11.8	100.0	73.33	9.8308
pctfrl	.0	70.7	28.03	13.8705
pctiep	.0	31.6	13.68	4.8574
pctlep	.0	8.4	1.28	2.0582
pctnonwhite	.0	38.3	9.18	6.9557
Nkids	19	1298	423.89	304.770

Note. necap=Grade 6 math prior achievement on NECAP assessment; gend=gender; male=1; iep=students with identified disabilities; frl= free- or reduced-price lunch; lep=limited English proficient; pctmathprof=school level percent of students who were proficient or above in math on NECAP assessment; pctfrl=school level percent of students who qualify for free- or reduced-price lunch; pctiep=school level percent of students with identified disabilities; pctlep=school level percent of students identified as limited English proficient; pctnonwhite=school level percent of non-White students; Nkids=number of students in the school who took SBAC.

Prior to centering the continuous predictor and control variables, a quick snapshot of achievement by year and treatment was analyzed, alongside descriptives for each of the PACE schools by treatment year. This information provides some context about unconditional treatment effects and treatment schools prior to the multivariate analyses.

Table 4.2 shows the unconditional mean Grade 8 SBAC math scale scores by school year and number of treatment years in the unweighted and weighted sample for comparison. Eighth grade students receiving one year of treatment in the 2015-16 school year had the highest unconditional mean Grade 8 math achievement in both samples; whereas the lowest unconditional mean Grade 8 math achievement was for students receiving one year of treatment in the 2014-15 school year in both samples.

Table 4.2 Unconditional mean Grade 8 math scale scores by year and treatment status in the unweighted and inverse propensity score weighted sample

		Unweighted Sample			Weighted Sample		
SBACyearid	#Treatyrs	N	Mean	SD	N	Mean	SD
1415	0	10335	2572.08	101.009	13148	2568.46	99.879
	1	477	2535.47	98.190	5942	2534.35	94.795
	pooled	10812	2570.46	101.161	19090	2557.84	99.583
1516	0	10064	2580.60	101.629	11147	2579.40	102.199
	1	300	2587.70	108.021	1977	2587.16	102.064
	2	456	2562.64	105.940	6013	2571.55	105.249
	pooled	10820	2580.04	102.058	19136	2577.74	103.258

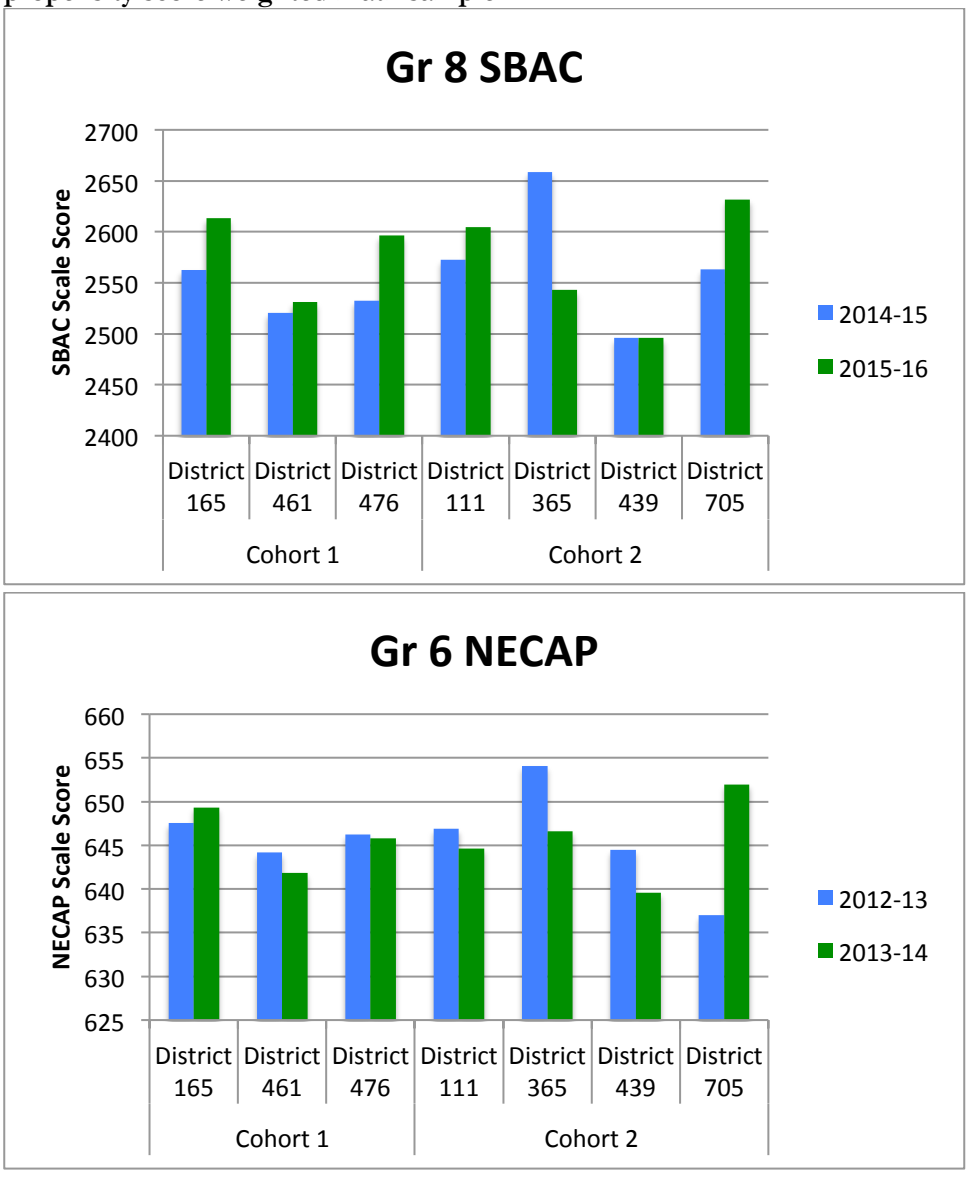
Descriptive statistics for the seven PACE schools/districts[16] by treatment year in the weighted and unweighted sample are provided in Appendix C. There are two main differences between PACE schools. First, as might be expected, PACE schools differ in their mean Grade 8 SBAC math performance over the two years of the pilot in a similar pattern to how they differed in their mean Grade 6 NECAP math performance. This isn't surprising given the fact that these are the same groups of students—this data is repeated cross-sectional not longitudinal.

Figure 4.1 illustrates this variability using a bar graph. In the top panel, the blue bars represent the 2014-15 school year and shows how the PACE schools/districts differ in their mean Grade 8 SBAC math performance by district and by cohort. The green bars represent the 2015-16 school year. In the bottom panel, variability in prior Grade 6 NECAP math achievement is shown. The blue bars represent the 2012-13 school year and correspond to the blue bars in the top panel because the data represents the same group of students. Similarly, the green bars represent the 2013-14 school year and correspond to the green bars in the top panel. Due to differences in scale scores it is difficult to make precise comparisons between the two panels, however, the pattern of bars for three districts (Districts 461, 476, and 111) are reversed in the top panel as compared to the bottom panel. For example, in District 476, Grade 6 students in 2012-13 and 2013-14 tended to perform

[16] Schools/districts is used synonymously because there is only one PACE school per district with Grade 8 students.

around the same on average, but the same two cohorts of students performed around 60% of a standard deviation different on Grade 8 SBAC with students in the 2015-16 outperforming the cohort before them.

Figure 4.1 Unconditional school mean Grade 8 SBAC math scale score (top panel) and Grade 6 NECAP math scale score (bottom panel) by PACE districts, cohort, and year using the inverse propensity score weighted math sample



A second noticeable difference between PACE schools is the size of the Grade 8 cohort. For example, in District 365 (School 20885) there are only 5 eighth-grade students in the entire school in 2015-16 and they happen to all be female. This is in comparison to District 461 (School 22705) that in the same school year had 279 eighth-grade students with an almost even split between males and females. Other details on PACE districts, and in particular, how each district implemented the PACE pilot was provided under study context in Chapter Three (see Table 3.5). Readers are also referred to a formative evaluation of PACE that includes Tier 1 district perceptions about PACE implementation gathered from site visits, classroom observations, teacher surveys, and focus group interviews with students, parents, teachers, and administrators during the 2016-17 school year (Becker et al., 2017).

Multi-Level Model Analyses

Table 4.3 presents parameter estimates and goodness of fit statistics from selected multi-level models fit. I began by fitting a fully unconditional model (**M0**), which allowed me to estimate the intraclass correlation coefficient—or the amount of variation in Grade 8 math achievement that occurs within-schools as opposed to between-schools. If there is a very small amount of between-school variance (e.g., <1%), multi-level modeling may not be necessary because there is not a lot of clustering between schools to explain. Second, I fit a model that contains level-1 predictors only and examines how much of the within-school variance in Grade 8 math achievement can be "explained" by the level-1 covariates in the model. The only random effect in this model is the intercept (**M1**). Third, I fit a means as outcomes model (**M2**). This model contains only level-2 predictors and examines how much of the between-school variance in Grade 8 math achievement can be "explained" by the level-2 covariates in the model. Fourth, I fit a model that combined the level-1 and level-2 models (**M3**). The final model builds from M3 and tests for cross-level effects between years of treatment and all level-1 covariates added one at a time (**M4**).

The purpose for fitting models in this way is that it allows a researcher to examine the effect of each variable on the pseudo-R^2 statistic (i.e., how much of the variance in Grade 8 math achievement can be explained by the inclusion of that variable) and model fit statistics. It also allows for non-significant effects to be removed at level-1, for example, so that the most parsimonious measurement model can be used in subsequent model building. A full taxonomy of models can be found in Appendix D. The MIXED procedure in SPSS using Maximum Likelihood estimation was used to estimate all models. The inverse propensity score weight was used as a regression weight in all models.

The sensitivity of treatment effects to weighting was examined for robustness by comparing treatment effects estimated from the weighted sample with the treatment effects estimated from the unweighted sample (see Appendix E). Evidence of selection bias was found because there are differences in average treatment effect estimates between the weighted and unweighted analyses. Also, to check the robustness of the effect estimates, I fit the outcome models for each individual cohort instead of grouping both cohorts together. The yearly results were similar to the ones with all cohorts together.

Table 4.3 Parameter estimates and goodness of fit statistics from selected multi-level models showing the effects of student- and school-level characteristics on Grade 8 math achievement for the inverse propensity score weighted sample

	M0: Null		M1: Level-1 Only		M2: Level-2 Only		M3: Levels 1&2		M4: Cross-Level	
Variables	B	SE	B	SE	B	SE	B	SE	B	SE
Intercept	2577.47***	3.56	2587.11***	2.15	2570.31***	2.71	2578.55***	2.19	2576.92***	2.25
necap			6.26***	0.04			6.35***	0.04	6.29***	0.05
frl			-12.49***	0.98			-12.79***	0.95	-14.24***	1.27
iep			-11.38***	1.38			-9.93***	1.34	-16.75***	1.71
male			-13.18***	0.85			-13.72***	0.83	-7.86***	1.02
pctmathprof					1.11***	0.09	-0.38***	0.06	-0.34***	0.06
pctfrl					-1.01***	0.15	-0.52***	0.12	-0.42***	0.12
Nkids					-0.02*	0.01	-0.02***	0.01	-0.02***	0.01
sbacid					11.14***	1.75	13.83***	1.09	13.91***	1.08
treat1					-32.86*	14.51	-30.44*	12.08	-25.42*	12.30
treat2					-7.80	14.52	-4.31	12.08	0.33	12.33
sbacid*treat1					43.44*	15.43	43.90***	12.56	44.53***	12.71
treat1*necap									-0.54***	0.10
treat2*necap									1.08***	0.12
treat1*frl									3.02	2.29
treat2*frl									3.16	2.53
treat1*iep									13.63***	3.26
treat2*iep									27.79***	3.78
treat1*male									-17.49***	2.11
treat2*male									-17.68***	2.30
Variance components										
σ^2	16605.63***	160.13	6659.02***	64.22	16295.03***	157.18	6277.09***	60.54	6195.46***	59.75
τ_{00}	1335.36***	199.78	449.39***	69.14	593.56***	103.38	417.14***	64.53	428.49***	66.25
%Reduction σ^2			0.60		0.02		0.62		0.63	
%Reduction τ_{00}			0.66		0.56		0.69		0.68	
Goodness of fit										
-2LL	267238.75		247453.74		266750.68		246174.46		245895.4	
AIC	267244.75		247467.74		266770.68		246202.46		245939.4	
BIC	267268.70		247523.62		266850.50		246314.20		246115.0	

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

Note. MIXED command in SPSS with ML estimation; ATE inverse propensity score weights applied as a regression weight. B=unstandardized parameter coefficient; SE=standard error; AIC = Akaike's Information Criteria; BIC = Schwarz's Bayesian Criterion.

Based on the unconditional model (**M0**), the average school-level Grade 8 math achievement for the weighted sample was approximately 2577 on the Smarter Balanced summative math assessment over the first two years of the pilot. The estimated population variance in intercepts *within* schools is significant ($\hat{\sigma}^2$= 16605.63, p<.001), which means that students within schools differ in their average Grade 8 math achievement. The estimated population variance in intercepts *between* schools is also significant ($\hat{\tau}_{00}$=1335.36, p<.001), which means that schools differ in their average Grade 8 math achievement. The intraclass correlation coefficient suggests that about 7% of the variance in Grade 8 math achievement is between schools and the other 93% is within schools. Between-school variance on achievement test scores is typically around 20-25% of the total variance (Hedges & Hedberg, 2007), which means this is a small ICC. The ICCs estimated in the unweighted sample is 11%, which is similar to the weighted sample (see Appendix E).

The model in Table 4.3 (**M1**) included four level-1 predictors: prior achievement (necap), IEP status (iep), FRL status (frl), and gender (male). The dummy variables for limited English proficient (lep) and non-White (nonwhite) were not included in the regression models because of the low percentage of students in the sample classified as limited English proficient (1%) or non-White (8%). Level-1 predictors were added one at a time as fixed effects. Prior achievement was grand mean centered because when it was group mean centered the within-school variance estimate went up in comparison to the unconditional model. The pseudo-R^2 statistic for Model 1 was approximately 60%, which is the amount of "explainable" parameter variance in Grade 8 math achievement within schools explained by the level-1 fixed effects and random effect (intercept) in the model in comparison to the unconditional model (M0). Controlling for prior achievement also explained about 65% of the variability between-schools.

The means as outcomes model (**M2**) accounted for about 56% of the "explainable" parameter variance in average Grade 8 math achievement between schools in comparison to the unconditional model. Model 2 included a dummy variable for SBAC year ID (sbacid) and four level-2 control variables all grand-mean centered using the weighted sample mean. The level-2 control variables included percentage of students in the school proficient or above on the NECAP math test (pctmathprof), percentage of students in the school who qualify for free- and reduced-price lunch (pctfrl), and the number of students in the school who took SBAC (Nkids). I also fit models that controlled for the percentage of students in the school who have an IEP (pctiep), but the between-school variance estimate went up and the fixed effect was non-significant (see taxonomies in Appendix D). The between-school variance may have gone up because there are some schools in the sample (including in the PACE group) that have no IEP students in the school[17]. The percentage of limited English proficient students in the school (pctlep) and percentage of non-White students in the school (pctnonwhite) were not included as control variables in regression models because the sample mean was around 1% and 9%, respectively.

Model 2 also included the two treatment variables that were used to answer the research questions (treat1= "one year of treatment" and treat2= "two years of treatment"). As noted previously, treatment status was modeled using two dummy variables to allow for non-linear effects and dosage effects to be modeled. An interaction between SBAC year ID and treat1 was included to examine whether treatment effects differ by treatment year. There is no interaction between treat2 and SBAC year ID because schools implementing PACE for two years only took the SBAC test in the 2015-16 school year.

[17] No IEP students in the school is an artifact of school size and not an issue with identifying students with learning disabilities. For example, school 20855 has only 5 students in Grade 8 in 2015-16 – all of whom are female and none of whom have been identified with a disability (see Appendix B).

Model 3 (**M3**) combines level-1 and level-2 predictors. Prior to controlling for any cross-level effects, there is a negative effect of one year of PACE treatment during the 2014-15 school year (β=-30.44, p <.05) and two years of treatment during the 2015-16 school year (β=-4.31, p >.05)—although it is not statistically significant. During the 2015-16 school year, there is a positive effect of one year of PACE treatment around 13-points (p <.001). Model 3 accounts for about 62% of the explainable within-school variance and 69% of the explainable between-school variance in Grade 8 math achievement.

Model 4 (**M4**), the final model, includes all significant level-1 and level-2 control variables and also tests for cross-level effects between the two treatment variables and level-1 covariates. There is about a 63% reduction in within-school variance and a 68% reduction in between-school variance for Model 4 in comparison to the fully unconditional model (M0). Model fit indices such as Akaike's Information Criteria (AIC) and Schwarz's Bayesian Criterion (BIC) suggest that Model 4 has the best model fit of any of the models and was subsequently be used to answer the research questions.

The following equation represents Model 4:

$$Y_{ij} = \beta_0 + \beta_1 \boldsymbol{X}_{ij} + \beta_2 \boldsymbol{S}_j + \beta_3(\text{treat1}_j) + \beta_4(\text{treat2}_j) + \beta_5(\text{sbacid}_j) + \beta_6(\text{treat1}_j * \text{sbacid}_j) + \beta_7(\text{treat1}_j * \boldsymbol{X}_{ij}) + \beta_8(\text{treat2}_j * \boldsymbol{X}_{ij}) + u_{0j} + r_{ij}$$

where Grade 8 math achievement of student *i* in school *j* (Y_{ij}) is a function of a vector of that student's observable characteristics (X_{ij}), school characteristics (S_j), treatment effects indicating either one or two years of dosage (treat1_j or treat2_j), SBAC year (sbacid_j), interactions between one year of treatment and SBAC year, interactions between treatment effects and observable student characteristics, the random effect of the intercept (u_{0j}), and a residual term that captures the random noise that may occur at the student-level (r_{ij}).

Findings for Research Question #1

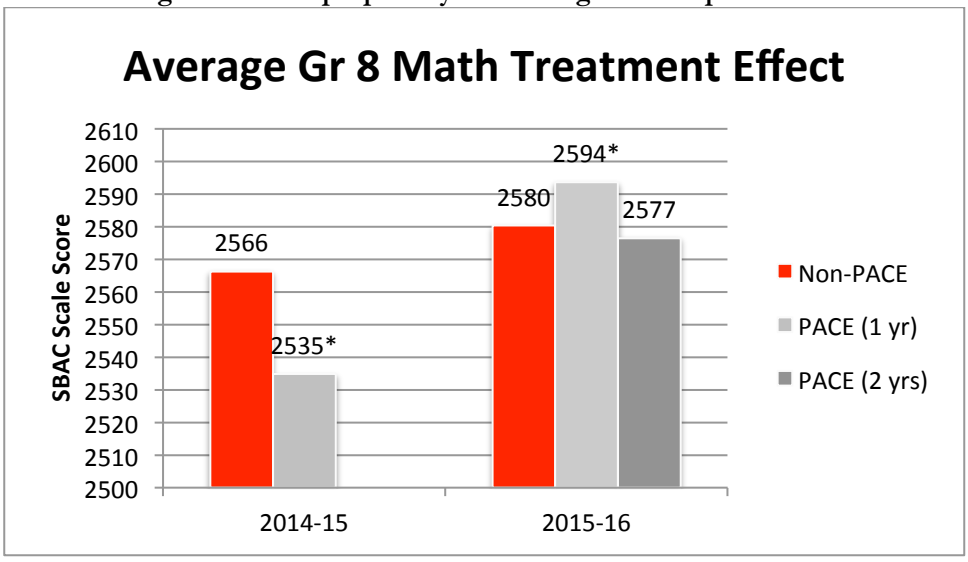
The first research question in math examines the average effect of the PACE pilot on Grade 8 math student achievement outcomes when comparing students with similar probabilities of being selected into treatment. The variables of interest are treat1 and treat2 along with the interactions between treat1 (or one year of treatment) and SBAC year ID (0 = "2014-15" or 1 = "2015-16). The interaction between treat2 (or two years of treatment) with the SBAC year ID cannot be tested because students receiving two years of treatment only took the SBAC in 2015-16. Because of the significant interaction between treat1 and SBAC year ID in Model 4, the main effect of treat1 varies according to the year the students took the SBAC test. When SBAC year ID is "0" or the 2014-15 school year, there is no interaction and the main effect of treat1 is not significantly different from zero on the Grade 8 SBAC math achievement test.

Figure 4.2 illustrates the predicted mean Grade 8 SBAC math scale score by school year and treatment status for the average student. Non-PACE students are coded in red. The two average treatment effects that differ significantly at the .05-alpha level between PACE and non-PACE comparison students is for Grade 8 students who received one year of treatment during either school year, as denoted by the asterisks. The conditional average treatment effect is negative in Year 1 of the pilot (treat1: β = -24.44, $p < .05$). In Year 2 of the pilot, the conditional average treatment effect is positive because of the interaction with SBAC year ID (sbacid*treat1: β = 44.83, $p < .01$). This means that PACE students tend to perform around 30-points lower on Grade 8 math SBAC in Year 1 of the pilot when all other variable values are set to their sample average (d=-0.30). Starting in Year 2 of the pilot, PACE students tend to perform around 14-points higher under the same conditions (d=0.14). There is also a positive effect of two years of treatment (treat2: β = 1.40, $p >$ .05), however, because the figure shows effects for the average student and there is a negative interaction between two years of treatment and gender, the average effect for the average student is

3-points lower for PACE students receiving two years of treatment in comparison to the non-PACE group.

In terms of the practical significance of these findings, it is important to note that a *p*-value is the estimated probability that a difference that large would be found when, in fact, there was no difference in the population from which the sample was drawn. However, I am using the population of 8th grade students in this analysis and therefore treatment effects are practically significant even if they are not statistically significant because they reflect the magnitude of the effect for the Grade 8 student population in NH.

Figure 4.2 Mean Grade 8 SBAC math scale score by school year and treatment status for the average student using the inverse propensity score weighted sample



Note. Statistically significant differences between treatment groups are marked with an asterisk. Non-significant treatment effects are included. Figure represents the average student. Covariates in the model include student-level characteristics (prior achievement, free-and-reduced price lunch status, disability status, and gender), school-level characteristics (percent of students in the school who are math proficient or above, percent of students in the school who qualify for free-and-reduced price lunch, and number of students in the school who took SBAC), year ID and treatment variables (SBAC Year ID, 1 or 2 years of treatment, interaction between SBAC Year ID and treatment variables), and cross-level interactions between treatment variables and three student-level characteristics (1 or 2 years of treatment interacted with prior achievement, disability status, and gender).

These findings have at least a couple of implications. First, there is evidence for lower average math performance for PACE students than non-PACE students in the first year of implementation. It is important to note that the pilot was not officially approved until March 2015, which is two-thirds of the way through the 2014-15 school year and likely about a month before students took the SBAC assessment. It is unclear, therefore, whether lower math performance for PACE students in Year 1 is an artifact of an implementation dip often associated with an innovation or relatively little treatment. Second, findings suggest a positive effect in math of PACE for students receiving one year of treatment starting in the second year of the PACE pilot and basically no effect in math of PACE for students receiving two years of treatment in the second year, on average. Overall, these findings suggest that PACE treatment has a different effect on Grade 8 math student achievement outcomes depending upon the year of the pilot and number of treatment years. More years of data would help to uncover the extent to which there are patterns of effects that remain consistent over time, especially as these are different cohorts of students.

Findings for Research Question #2

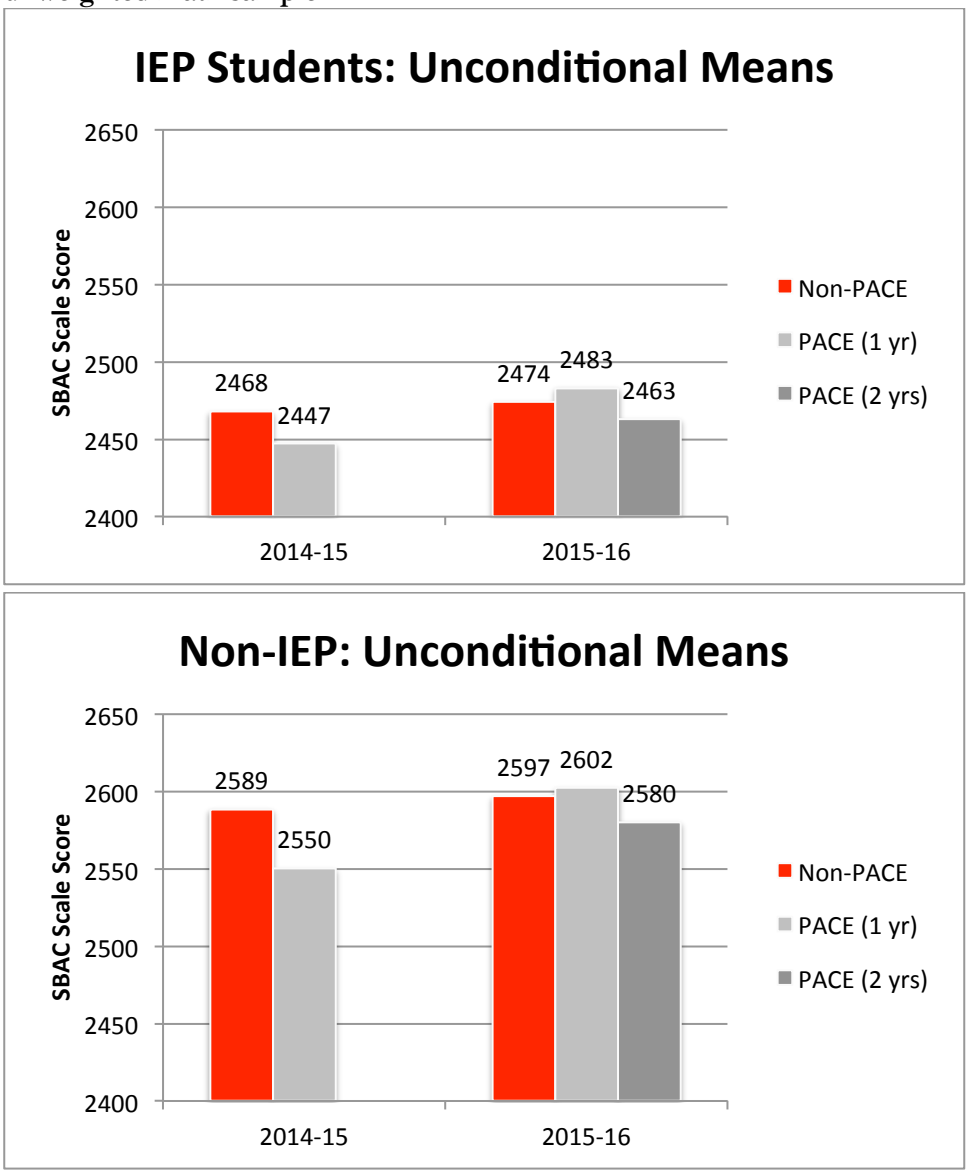
The second research question in math investigates whether the average treatment effect for PACE vs. non-PACE comparison students varies based on different student-level characteristics such as prior achievement, free- and reduced-price lunch, disability status, or gender. As such, the variables of interest in this investigation are the cross-level effects between treat1 and treat2 with the level-1 main effects associated with prior achievement (necap), IEP status (iep), and gender (male). The cross-level effect with free- and reduced-price lunch (frl) was also examined, however, it was non-significant. The findings below are organized by level-1 variable and use parameter estimates from Model 4 above.

Prior achievement. In general, there is a relatively small interaction effect between student prior achievement and treatment variables (treat1*necap: β = -0.55, p <.001; treat2*necap: β = 1.05, p <.001) that moderates the effect of prior achievement on outcome for PACE students. Since the main effect of student prior achievement on Grade 8 math achievement is positive (β = 6.30, p <.001), this means that Grade 8 students who received one year of treatment tend to exhibit a slightly smaller positive effect of prior achievement on Grade 8 math achievement, on average; whereas students who received two years of treatment tend to exhibit a slightly larger positive effect of prior achievement on Grade 8 math achievement, on average. Overall, these findings suggest that there is a small differential effect of student prior achievement for some PACE students that is negative for students with one year of treatment, but positive for students with two years of treatment.

Disability status. In examining differential effects by disability status, a few really interesting findings emerged. First, PACE students on IEP plans tend to exhibit higher average Grade 8 math achievement in comparison to their non-PACE comparison peers who also have IEP plans starting in Year 2, holding all other predictors constant. This is because there were statistically significant positive interactions between disability status and treatment. However, simply computing effects for the average IEP and non-IEP student by treatment status using the parameter estimates from the final model (Model 4) makes it appear as if there is no longer any achievement gap between PACE IEP and PACE non-IEP students in Grade 8 math. And yet from examining the unconditional mean SBAC scores for students with and without disabilities by treatment status using the unweighted sample to get a sense of the average observed achievement without any student- or school-level controls and without weighting, there is still an achievement gap between IEP and non-IEP students in the PACE group (see Figure 4.3). Similar findings were noted when re-examining these analyses using the inverse propensity score weighted sample. In other words, students with

disabilities in both groups (PACE and non-PACE) still perform lower, on average, in comparison to students without disabilities in both groups.

Figure 4.3 Unconditional Grade 8 SBAC math scale scores for IEP and non-IEP students using the unweighted math sample



To investigate what might be driving the positive interaction effects for PACE IEP students, I first examined frequency counts to get a sense of how many PACE students had IEPs. There are 230 PACE students with IEPs in the unweighted sample (14% of the PACE group) and over half (44%) of those students are from one school (22705). Next, I examined whether the interaction

effects might be an artifact of the weighting applied, but similar interaction effects were found using the unweighted sample (see Appendix E). Also, because interaction effects can be an artifact of outliers, the analysis was rerun without the most extreme cases (5 highest and 5 lowest student-level residuals[18]) for both PACE and non-PACE groups (i.e., 20 students total were removed) and the results were replicated.

In order to examine IEP effects by school to see if certain influential schools were driving the positive interaction effects, I fit separate regressions for each school in the analytic sample using student-level Grade 8 SBAC math scale scores as the outcome variable. Covariates included prior achievement, free- and reduced-price lunch status, gender, and disability status. One of the 7 PACE schools had no IEP students in either year (20885) and therefore had no parameter associated with the effect of disability status on Grade 8 student achievement outcomes in math. One PACE school only had IEP students in the second year (26505) and another only had IEP students in the first year (28400)—but it was not implementing PACE in that year. This left five of the seven PACE schools with effects of IEP status on outcome that could be examined. Of the five PACE schools with IEP students in one or both years, three had positive effects of IEP status (22705, 26550, 26505)(N=123; 53% of PACE IEP students) and two had negative effects of IEP status (20270, 20630)(N=106; 46% of PACE IEP students).

In trying to ascertain whether there were influential cases driving these positive effects in the two PACE schools where a positive effect of IEP was exhibited (22705, 26550), I examined the student-level residuals resulting from the final multi-level model specification for IEP students and non-IEP students attending these PACE schools. It appears that School 22705 is driving the positive effects for two reasons. First, School 22705 has the largest IEP student population of all

[18] I used the student-level residuals rather than the 5 highest and 5 lowest SBAC scale scores because there were many students with the exact same score at the top and bottom of the score distribution. Removing 20 high and low performing students would then be based on an arbitrary decision.

PACE schools with 52 IEP students in Year 1 and 50 IEP students in Year 2 (44% of the total PACE IEP population). Second, the mean residuals for IEP students are more positive than the mean residuals for non-IEP students in School 22705. For example, in Year 1, the mean residual for IEP students in School 22705 was about 29 points (SD=73.4; Min=-145.18; Max=196.49). In Year 2, the mean residual for IEP students was around 8 points (SD=75.5; Min=-152.18; Max=197.01). In those same years, non-IEP students attending School 22705 had mean residuals of 1.6 and -8.4 points, respectively.

It appears, therefore, that there are two factors contributing to the positive interaction effects between IEP status and treat1 and treat2. First, a larger percentage of PACE IEP students exhibited positive effects of disability status on the outcome. Second, in the schools with positive effects of IEP status, IEP students performed better than expected in comparison to non-IEP students.

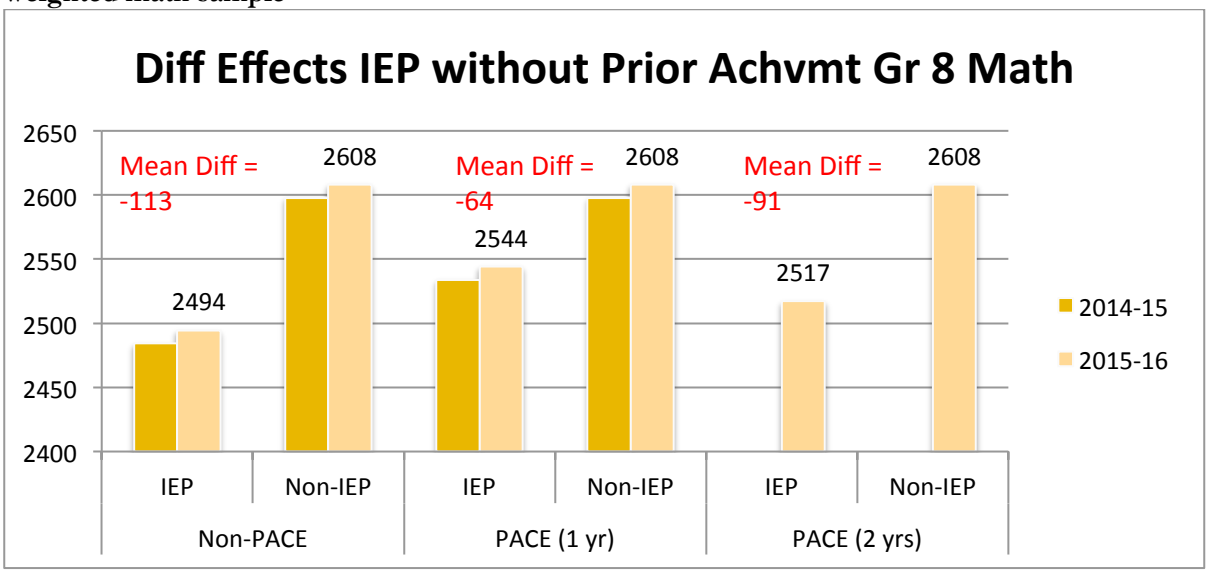
Overall, these findings suggest that (in general) there is a positive differential effect for PACE students with identified disabilities in comparison to their non-PACE peers who have also been diagnosed with a disability. These findings also suggest that IEP students from two PACE schools are largely driving that positive effect. However, these analyses do not explain why there appears to be no achievement gap between PACE IEP and non-IEP students—a finding which does not seem likely given the unconditional mean SBAC scale scores for these two groups of PACE students.

Instead, further analyses suggest that the appearance of no achievement gap for PACE IEP vs. PACE non-IEP students is an artifact of controlling for prior achievement in the model. Controlling for prior achievement means that only students of similar prior achievement are being compared. The question becomes: Is this a fair comparison for IEP students who likely differ widely from one another for many reasons including disability category (e.g., speech/language impairment,

intellectual disability, emotional disability, hearing impairment, autism, etc.), which is also related to prior achievement levels? For example, what is the likelihood that a student with a hearing impairment who attends a PACE school who demonstrated high prior math achievement would perform worse than a PACE non-IEP student who doesn't have a hearing impairment, but also has high prior achievement?

In order to isolate the differential effect of treatment for students with disabilities on Grade 8 math achievement, I re-fit the final model without prior achievement as a student-level control variable. The parameter estimate associated with the effect of IEP status on Grade 8 math achievement is sizable (β = -113.4, p <.001) and the interactions between IEP status and treatment are still positive (treat1*iep: β = 49.6, p <.001; treat2*iep: β = 22.6, p <.001). Figure 4.4 visually depicts the effects of IEP status on Grade 8 math achievement and shows how there is a narrowing of the achievement gap for PACE IEP students such that the mean difference between IEP and non-IEP students shrinks from -113 points to -64 or -91 points, but there is still an achievement gap.

Figure 4.4 Differential effects of IEP status on Grade 8 SBAC math achievement using parameter estimates from a model that does not control for prior achievement for the inverse propensity score weighted math sample



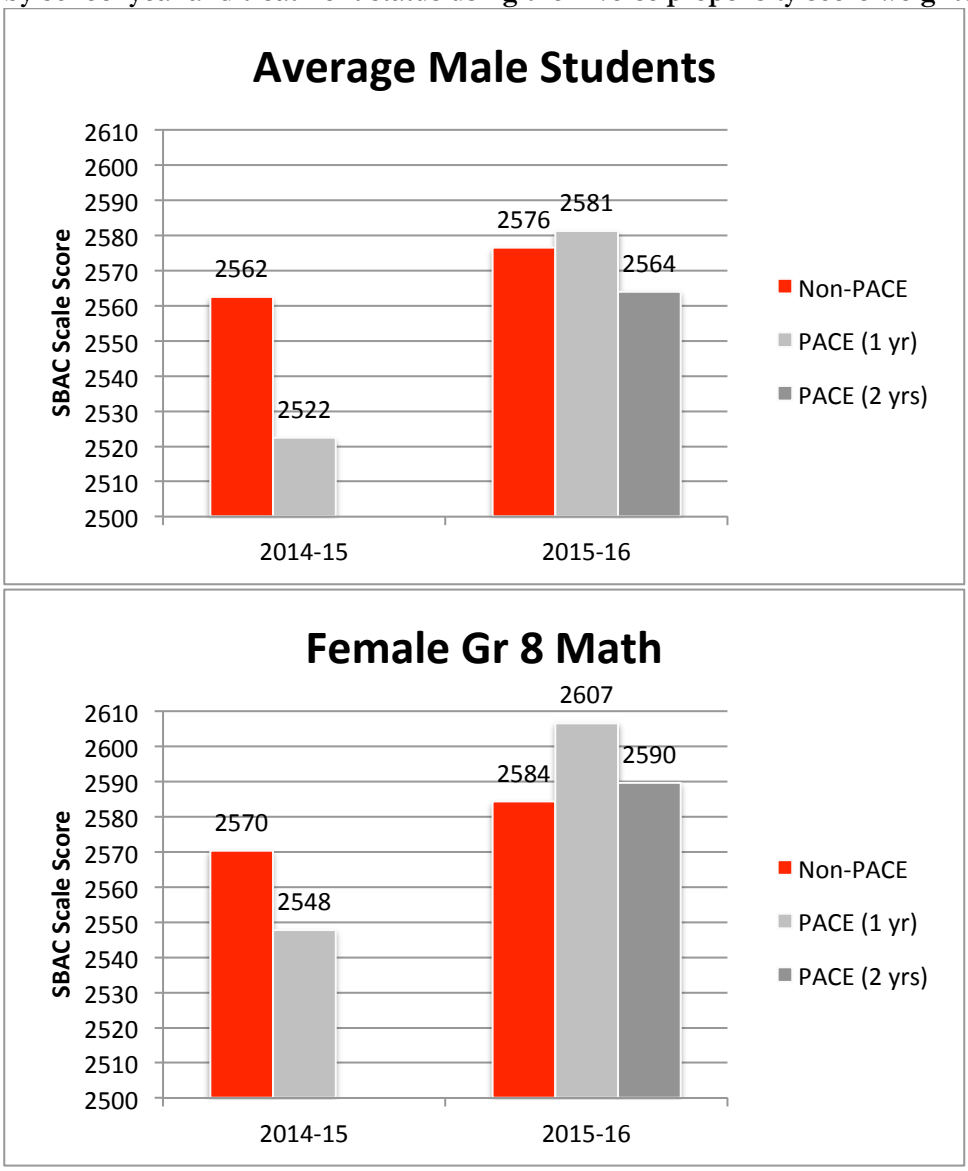
Note. Covariates in the model include student-level characteristics (free-and-reduced price lunch status, disability status, and gender), school-level characteristics (percent of students in the school who are math proficient or above, percent of students in the school who qualify for free-and-reduced price lunch, and number of students in the school who took SBAC), year ID and treatment variables (SBAC Year ID, 1 or 2 years of treatment, interaction between SBAC Year ID and treatment variables), and cross-level interactions between treatment variables and two student-level characteristics (1 or 2 years of treatment interacted with disability status and gender).

This finding is significant because narrowing the achievement gap by 20 to 50 points in comparison to non-PACE IEP students (about 20-50% of the pooled SBAC standard deviation), which is substantial. Given that those positive effects are likely driven by two PACE schools (22705 and 26550) and there is a small number of PACE IEP students in the sample, these findings should be considered exploratory and in need of replication. Future research could examine differential effects for students with disabilities in other grade levels, with a larger sample size, and using different methods. Future research could also investigate the extent to which these two PACE schools employ different special education models and processes or have different populations of special education students. Future research could also examine differences in treatment effects for students in different disability categories, although that information is not currently available from the state.

Gender. Figure 4.5 shows the mean Grade 8 SBAC math scale score for male students in the top panel and female students in the bottom panel for the average student. Overall, there are two main findings I want to highlight. First, female students tend to outperform male students in similar years of the pilot and treatment status. This is because the main effect of gender (male=1; female=0) is negative as are the interactions between gender and treatment status. Second, as a result of the negative interactions between treatment status and gender, male PACE students tend to perform about the same as their male non-PACE counterparts in the second year of the pilot.

It is unclear why male students in NH tend to not perform as well on the Grade 8 math assessment in comparison to female students. Nationally, male and female students in Grade 8 math tend to perform around the same, on average (National Center for Education Statistics, 2017). It is also not clear why there is a negative interaction between gender and treatment status. Prior research on performance assessment programs and competency-based education did not examine differences in effects by gender, so it is unclear whether this is a common pattern or not. This could be an area of future research, especially the extent to which this pattern holds over time and in other grade/subject combinations.

Figure 4.5 Mean Grade 8 SBAC math scale scores for males (top panel) and females (bottom panel) by school year and treatment status using the inverse propensity score weighted sample



Note. Figure represents the average male or female student. Non-significant treatment effects are included. Covariates in the model include student-level characteristics (prior achievement, free-and-reduced price lunch status, disability status, and gender), school-level characteristics (percent of students in the school who are math proficient or above, percent of students in the school who qualify for free-and-reduced price lunch, and number of students in the school who took SBAC), year ID and treatment variables (SBAC Year ID, 1 or 2 years of treatment, interaction between SBAC Year ID and treatment variables), and cross-level interactions between treatment variables and three student-level characteristics (1 or 2 years of treatment interacted with prior achievement, disability status, and gender).

Findings for Research Question #3

In order to examine how average treatment effects vary among PACE schools, I used the *lme4* package in R (Bates et al., 2016) to obtain the level-2 residuals by School ID. I am interested in the extent to which PACE schools performed better than predicted (positive residuals) or worse than predicted (negative residuals) and if there are any patterns across PACE schools or pilot years. In other words, are PACE schools performing better than expected or worse than expected, based on the level-2 residuals from Model 4? Table 4.4 summarizes the descriptive statistics for the level-2 residuals.

Table 4.4 Descriptive statistics on level-2 residuals for Grade 8 math using the inverse propensity score weighted sample

Mean	0.00
Median	-1.55
Std. Deviation	19.31
Range	114.04
Minimum	-56.50
Maximum	57.53

Note. Covariates in the model include student-level characteristics (prior achievement, free-and-reduced price lunch status, disability status, and gender), school-level characteristics (percent of students in the school who are math proficient or above, percent of students in the school who qualify for free-and-reduced price lunch, and number of students in the school who took SBAC), year ID and treatment variables (SBAC Year ID, 1 or 2 years of treatment, interaction between SBAC Year ID and treatment variables), and cross-level interactions between treatment variables and three student-level characteristics (1 or 2 years of treatment interacted with prior achievement, disability status, and gender).

Table 4.5 shows the level-2 residual values for PACE schools by year. Most PACE schools performed lower than predicted as indicated by the red font; however, PACE schools that participated in the pilot for two years tended to exhibit positive residuals in the second year of implementation (e.g., School 20630 and 26505). That said, with only two years of data for three schools it is not possible to make any claims about trends.

Table 4.5 Level-2 residuals for PACE schools by year for Grade 8 math using the inverse propensity score weighted sample

School ID	Level-2 Residuals Year 1	Level-2 Residuals Year 2
20270	*	-3.46
20630	-39.26	10.40
20885	*	-52.87
22705	-51.94	-55.52
26505	-22.61	14.20
26550	*	-38.59
28400	*	62.53

Note: *=Not yet implementing. Covariates in the model include student-level characteristics (prior achievement, free-and-reduced price lunch status, disability status, and gender), school-level characteristics (percent of students in the school who are math proficient or above, percent of students in the school who qualify for free-and-reduced price lunch, and number of students in the school who took SBAC), year ID and treatment variables (SBAC Year ID, 1 or 2 years of treatment, interaction between SBAC Year ID and treatment variables), and cross-level interactions between treatment variables and three student-level characteristics (1 or 2 years of treatment interacted with prior achievement, disability status, and gender).

Future research could investigate whether these patterns for some schools hold over time and the extent to which they can be explained by contextual within-district implementation factors. There is no apparent relationship between the school-level residuals and fidelity-of-implementation continuum verbalized by the NHDOE (Table 3.5).

English Language Arts/Literacy

Descriptive Analyses

To address the research question about the extent to which PACE students differ from non-PACE students with similar probabilities of being selected into treatment in terms of their student achievement outcomes in ELA, the first part of the analyses focused on descriptive and exploratory analyses. Descriptive statistics and distributions of all the variables were examined (see Table 4.6).

Table 4.6 Descriptive statistics on variables in the inverse propensity score weighted Grade 8 ELA sample (wtd. N_students=38,210)

	Minimum	Maximum	Mean	SD
necap	600	680	645.87	11.712
gend	0	1	.51	.500
iep	0	1	.14	.346
frl	0	1	.30	.459
lep	0	1	.01	.078
nonwhite	0	1	.08	.26341
pctmathprof	11.8	100.0	73.33	9.8308
pctfrl	.0	70.7	28.03	13.8705
pctiep	.0	31.6	13.68	4.8574
pctlep	.0	8.4	1.28	2.0582
pctnonwhite	.0	38.3	9.18	6.9557
Nkids	19	1298	423.89	304.770

Note. necap=Grade 6 math prior achievement on NECAP assessment; gend=gender; referrant group is male; iep=students with identified disabilities; frl= free- or reduced-price lunch; lep=limited English proficient; pctmathprof=school level percent of students who were proficient or above in math on NECAP assessment; pctfrl=school level percent of students who qualify for free- or reduced-price lunch; pctiep=school level percent of students with identified disabilities; pctlep=school level percent of students identified as limited English proficient; pctnonwhite=school level percent of non-White students; Nkids=number of students in the school who took SBAC.

Prior to centering the continuous predictor and control variables, a quick snapshot of achievement by year and treatment was analyzed, alongside descriptives for each of the PACE schools by treatment year. This information provides some context about unconditional treatment effects and treatment schools prior to the multivariate analyses. Table 4.7 shows the unconditional mean Grade 8 SBAC ELA scale scores by year and treatment status (number of treatment years) in the weighted and unweighted sample for comparison. In contrast to the math analyses, non-PACE

students in the 2015-16 school year had the highest unconditional mean Grade 8 ELA achievement in both the unweighted and weight sample. In both the weighted and unweighted samples, the lowest unconditional mean Grade 8 ELA achievement was for students receiving one year of treatment in the 2014-15 school year.

Table 4.7 Unconditional mean Grade 8 SBAC ELA scale scores by year and treatment status in the unweighted and inverse propensity score weighted sample

		Unweighted Sample			Weighted Sample		
SBACyearid	#Treatyrs	N	Mean	SD	N	Mean	SD
1415	0	10224	2586.05	85.218	13010	2585.19	84.919
	1	461	2557.56	90.156	5798	2554.99	87.031
	pooled	10685	2584.82	85.628	18808	2575.88	86.703
1516	0	9965	2595.78	86.966	11037	2594.28	87.497
	1	299	2587.80	91.868	1970	2589.45	89.684
	2	452	2578.17	86.493	5994	2589.24	89.924
	pooled	10716	2594.81	87.158	19002	2592.19	88.527

Descriptive statistics for the seven PACE schools/districts[19] by treatment year in the weighted and unweighted ELA sample are provided in Appendix F. Similar to the math analyses, PACE districts have different mean Grade 8 SBAC ELA performance over the two years of the pilot. These differences tend to mirror differences in Grade 6 NECAP ELA performance for the same student cohorts. Also similar to the math analyses, there are a few PACE districts with no IEP students in one or both pilot years and there is wide variability in the number of students within each school. Other details on PACE districts, and in particular, how each district implemented the PACE pilot was provided under study context in Chapter Three (see Table 3.5). Readers are also referred to the formative evaluation of PACE (Becker et al., 2017).

[19] Schools/districts is used synonymously because there is only one PACE school per district with Grade 8 students.

Multi-Level Model Analyses

Table 4.8 presents parameter estimates and goodness of fit statistics from selected multi-level models fit. I fit models using the same process and reasoning explained in the math section. A full taxonomy of ELA models can be found in Appendix G. The MIXED procedure in SPSS using Maximum Likelihood estimation was used to estimate all models. The inverse propensity score weight was used as a regression weight in all models. The sensitivity of treatment effects to weighting was examined for robustness by comparing treatment effects estimated from the weighted sample with the treatment effects estimated from the unweighted sample (see Appendix H). Evidence of selection bias was found because there are differences in average treatment effect estimates from the weighted and unweighted analyses. Also, to check the robustness of the effect estimates, I fit the outcome models for each individual cohort instead of grouping both cohorts together. The yearly results were similar to the ones with all cohorts together.

Table 4.8 Parameter estimates and goodness of fit statistics from selected multi-level models showing the effects of student- and school-level characteristics on Grade 8 ELA achievement for the inverse propensity score weighted sample

	M0: Null		M1: Level-1 Only		M2: Level-2 Only		M3: Levels 1&2		M4: Cross-Level	
Variables	B	SE	B	SE	B	SE	B	SE	B	SE
Intercept	2592.54***	3.02	2605.36***	2.30	2583.85***	2.09	2597.76***	2.15	2596.94***	2.20
necap			4.26***	0.04			4.24***	0.04	4.18***	0.05
frl			-11.01***	0.97			-11.45***	0.95	-14.39***	1.27
iep			-27.33***	1.38			-27.39***	1.35	-30.84***	1.71
male			-16.42***	0.86			-17.11***	0.85	-12.89***	1.05
pctELAprof					1.38***	0.09	0.39***	0.07	0.38***	0.07
pctfrl					-0.97***	0.12	-0.58***	0.11	-0.56***	0.12
Nkids					-0.03***	0.01	-0.04***	0.01	-0.04***	0.01
sbacid					10.55***	1.51	10.41***	1.09	10.33***	1.08
treat1					-30.01*	10.79	-33.35*	11.71	-35.80***	11.84
treat2					-7.28	10.79	-9.44	11.71	-0.57	11.87
sbacid*treat1					18.96	11.65	33.21*	12.20	32.71*	12.25
treat1*necap									0.15	0.10
treat2*necap									0.17	0.11
treat1*frl									10.75***	2.32
treat2*frl									1.58	2.50
treat1*iep									2.15	3.32
treat2*iep									18.28***	3.87
treat1*male									-2.87	2.17
treat2*male									-22.80***	2.36
Variance components										
σ^2	12527.89***	121.45	6483.21***	62.85	12220.53***	118.50	6224.13***	60.35	6183.17***	59.96
τ_{00}	956.79***	142.25	524.63***	77.89	324.46***	59.10	391.89***	61.56	395.17***	62.01
%Reduction σ^2			0.48		0.02		0.50		0.51	
%Reduction τ_{00}			0.45		0.66		0.59		0.59	
Goodness of fit										
-2LL	258371.12		244279.48		257736.45		243380.60		243240.83	
AIC	258377.12		244293.48		257756.45		243408.60		243284.83	
BIC	258371.12		244349.28		257836.17		243520.20		243460.19	

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

Note. MIXED command in SPSS with ML estimation; ATE inverse propensity score weights applied as a regression weight. B=unstandardized parameter coefficient; SE=standard error; AIC = Akaike's Information Criteria; BIC = Schwarz's Bayesian Criterion.

Based on the unconditional model (**M0**), the average school-level Grade 8 ELA achievement for the weighted sample was approximately 2593 on the Smarter Balanced summative ELA assessment over the first two years of the pilot. The estimated population variance in intercepts *within* schools is significant ($\hat{\sigma}^2$= 12527.89, p<.001), which means that students within schools differ in their average Grade 8 ELA achievement. The estimated population variance in intercepts *between* schools is also significant ($\hat{\tau}_{00}$=956.79, p<.001), which means that schools differ in their average Grade 8 ELA achievement. The intraclass correlation coefficient suggests that about 7% of the variance in Grade 8 ELA achievement is between schools and the other 93% is within schools. Between-school variance on achievement test scores is typically around 20-25% of the total variance (Hedges & Hedberg, 2007), which means this is a small ICC. The ICCs estimated in the unweighted ELA sample is 11%, which is similar to the weighted sample (see Appendix H).

The model (**M1**) included four level-1 predictors: prior achievement (necap), IEP status (iep), FRL status (frl), and gender (male). Nonwhite was also examined, but was not significant and therefore removed prior to subsequent modeling. LEP was not included due to the small percentage of students classified as limited English proficient in this sample (1%). Level-1 predictors were added one at a time as fixed effects. Only the intercept was modeled as a random effect. Prior achievement was grand mean centered because when it was group mean centered the within-school variance estimate went up in comparison to the unconditional model. The pseudo-R^2 statistic for Model 1 was approximately 48%, which is the amount of "explainable" parameter variance in Grade 8 ELA achievement within schools explained by the level-1 fixed effects and random effect (intercept) in the model in comparison to the unconditional model (M0). Controlling for prior achievement also explained about 46% of the variability between-schools.

The means as outcomes model (**M2**) accounted for about 66% of the "explainable" parameter variance in average Grade 8 ELA achievement between schools in comparison to the

unconditional model. Model 2 included a dummy variable for SBAC year ID (sbacid) and four level-2 control variables all grand-mean centered using the weighted sample mean. The level-2 control variables included percentage of students in the school proficient or above on the NECAP ELA test (pctELAprof), percentage of students in the school who qualify for free- and reduced-price lunch (pctfrl), and the number of students in the school who took SBAC (Nkids). Similar to the math analyses, I also fit models that controlled for the percentage of students in the school who have an IEP (pctiep), but the between-school variance estimate went up and the fixed effect was non-significant (see taxonomies in Appendix G). The between-school variance may have gone up because there are some schools in the sample (including in the PACE group) that have no IEP students in the school. Also similar to the math analyses, percentage of limited English proficient students in the school and percentage of non-White students in the school were not included as control variables in the regression models because the sample means were small—1% and 9%, respectively.

Model 2 also included the two treatment variables that were used to answer the research questions (treat1= "one year of treatment" and treat2= "two years of treatment"). As noted previously, treatment status was modeled using two dummy variables to allow for non-linear effects and dosage effects to be modeled. An interaction between SBAC year ID and treat1 was included to examine whether treatment effects differ by treatment year. There is no interaction between treat2 and SBAC year ID because schools implementing PACE for two years only took the SBAC test in the 2015-16 school year.

Model 3 (**M3**) combines level-1 and level-2 predictors. Prior to controlling for any cross-level effects, there is a negative effect of one year of PACE treatment during the 2014-15 school year (β=-33.35, p <.05) and two years of treatment during the 2015-16 school year (β=-9.44, p >.05)—although it is not statistically significant. During the 2015-16 school year, there is basically

no effect of one year of PACE treatment because of the interaction effect between SBAC year ID and treat1 (p <.05). Model 3 accounts for about 50% of the explainable within-school variance and 59% of the explainable between-school variance in Grade 8 ELA achievement.

Model 4 (**M4**), the final model, includes all significant level-1 and level-2 control variables and also tests for cross-level effects between the two treatment variables and level-1 covariates. There is about a 51% reduction in within-school variance and a 59% reduction in between-school variance for Model 4 in comparison to the fully unconditional model (M0). Model fit indices such as Akaike's Information Criteria (AIC) and Schwarz's Bayesian Criterion (BIC) suggest that Model 4 has the best model fit of any of the models and was subsequently be used to answer the research questions.

The following equation represents Model 4:

$$Y_{ij} = \beta_0 + \beta_1 \boldsymbol{X}_{ij} + \beta_2 \boldsymbol{S}_j + \beta_3(\text{treat1}_j) + \beta_4(\text{treat2}_j) + \beta_5(\text{sbacid}_j) + \beta_6(\text{treat1}_j * \text{sbacid}_j) + \beta_7(\text{treat1}_j * \boldsymbol{X}_{ij}) + \beta_8(\text{treat2}_j * \boldsymbol{X}_{ij}) + u_{0j} + r_{ij}$$

where Grade 8 ELA achievement of student *i* in school *j* (Y_{ij}) is a function of a vector of that student's observable characteristics (X_{ij}), school characteristics (S_j), treatment effects indicating either one or two years of dosage (treat1_j or treat2_j), SBAC year (sbacid_j), interactions between one year of treatment and SBAC year[20], interactions between treatment effects and student observable characteristics, the random effect of the intercept (u_{0j}), and a residual term that captures the random noise that may occur at the student-level (r_{ij}).

[20] There can be no interaction between two years of treatment (treat2) and SBAC year because students receiving two years of treatment only took SBAC in 2015-16.

Findings for Research Question #1

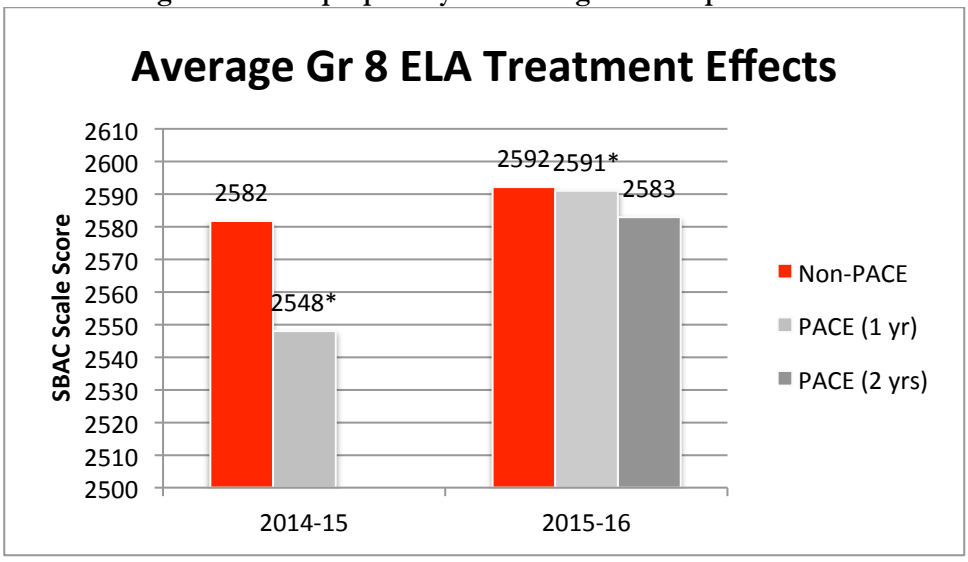
The first research question examines the average effect of the PACE pilot on Grade 8 ELA student achievement outcomes when comparing students with similar probabilities of being selected into treatment. Similar to the math analyses, the variables of interest are treat1 and treat2 along with the interactions between treat1 (or one year of treatment) and SBAC year ID (0 = "2014-15" or 1 = "2015-16). Because of the significant interaction between treat1 and SBAC year ID in Model 4, the main effect of treat1 varies according to the year the students took the SBAC test. When SBAC year ID is "0" or the 2014-15 school year, there is no interaction and the main effect of treat1 is not significantly different from zero on the Grade 8 SBAC ELA achievement test.

Figure 4.6 illustrates the predicted mean Grade 8 SBAC ELA scale score by school year and treatment status for the average student. Non-PACE students are coded in red. The two average treatment effects that differ significantly at the .05-alpha level between PACE and non-PACE comparison students is for Grade 8 students who received one year of treatment during either school year, as denoted by the asterisks. The conditional average treatment effect for Grade 8 ELA is negative in Year 1 of the pilot (treat1: β = -34.94, $p < .001$). In Year 2 of the pilot, there is almost no conditional average treatment effect because the positive interaction between SBAC year ID and one year of treatment almost cancels out the negative effect of treat1 (sbacid*treat1: β = 32.62, $p < .05$). This means that PACE students tend to perform around 34-points lower on Grade 8 ELA SBAC in Year 1 of the pilot when all other variable values are set to their sample average (d=-0.34). Starting in Year 2 of the pilot, PACE students tend to perform around the same as their non-PACE comparison peers under the same conditions. There is also a very small positive effect of two years of treatment (treat2: β = 0.38, $p > .05$), however, because the figure shows effects for the average student and there is a negative interaction between two years of treatment and gender, the

average effect for the average student is 9-points lower for PACE students receiving two years of treatment in comparison to the non-PACE group.

As in the math analyses, it is important to note that a *p*-value is the estimated probability that a difference that large would be found when, in fact, there was no difference in the population from which the sample was drawn. However, I am using the population of 8th grade students in this analysis and therefore treatment effects are practically significant even if they are not statistically significant because they reflect the magnitude of the effect for the Grade 8 student population in NH.

Figure 4.6 Mean Grade 8 ELA SBAC scale score by school year and treatment status for the average student using the inverse propensity score weighted sample



Note. Statistically significant differences between treatment groups are marked with an asterisk. Non-significant treatment effects are included. Figure represents the average student. Covariates in the model include student-level characteristics (prior achievement, free-and-reduced price lunch status, disability status, and gender), school-level characteristics (percent of students in the school who are math proficient or above, percent of students in the school who qualify for free-and-reduced price lunch, and number of students in the school who took SBAC), year ID and treatment variables (SBAC Year ID, 1 or 2 years of treatment, interaction between SBAC Year ID and treatment variables), and cross-level interactions between treatment variables and three student-level characteristics (1 or 2 years of treatment interacted with prior achievement, disability status, and gender).

Similar to the math findings, there is evidence for lower ELA performance for PACE students than non-PACE students in the first year of implementation. However, it unclear whether those effects are an artifact of an implementation dip or no treatment as the pilot was not officially approved until about a month before students took SBAC in the 2014-15 school year. In contrast to the math findings, PACE students are predicted to perform about the same or a little lower, on average, as their non-PACE comparison peers in ELA starting in Year 2. More years of data and analyses in other grades and subject areas would help elucidate the extent to which these findings of basically "no effect" of PACE on Grade 8 ELA achievement holds over time and in different grades, especially as these are different cohorts of 8th graders. Overall, these findings suggest that PACE students are provided an equitable opportunity to learn in Grade 8 ELA and they are not negatively impacted by PACE treatment.

Findings for Research Question #2

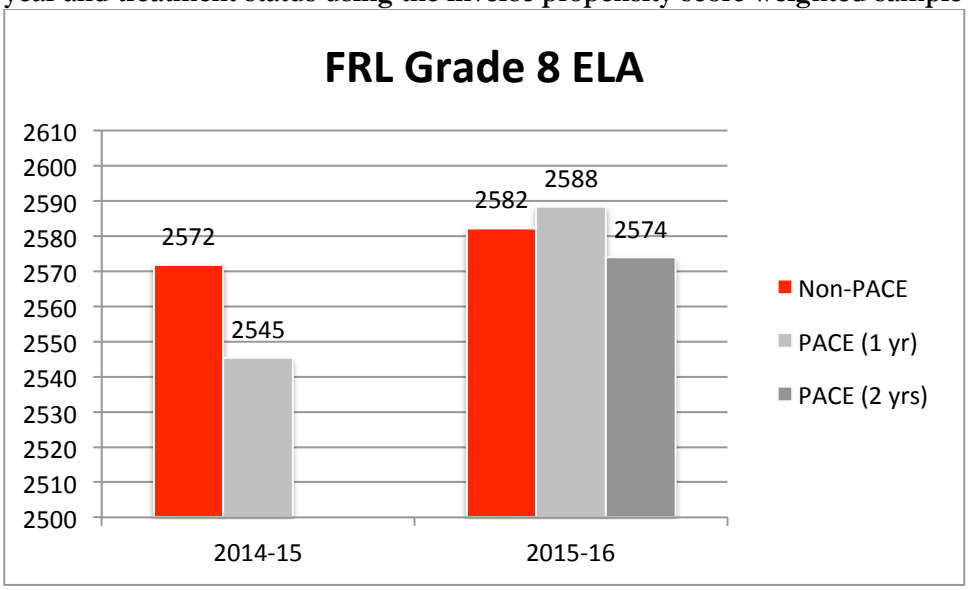
The second research question investigates whether the average treatment effect for PACE vs. non-PACE comparison students in Grade 8 ELA varies based on different student-level characteristics such as prior achievement, socioeconomic status, disability status, or gender. As such, the variables of interest in this investigation are the cross-level effects between treat1 and treat2 with the level-1 main effects associated with free- and reduced-price lunch (frl), IEP status (iep), and gender (male). The cross-level effect with prior achievement (necap) was also examined, however, it was non-significant. The findings below are organized by level-1 variable and use parameter estimates from Model 4 above.

Socioeconomic status (FRL). There is a significant positive interaction effect between free- and reduced-price lunch (FRL) status and one year of treatment, which means that the effect of FRL status on Grade 8 ELA achievement varies as a function of treatment. For example, although the main effect of FRL is negative (frl: β = -14.15, $p < .001$), for PACE students, that negative main

effect is lessened for students receiving one year of treatment because of the positive interaction with FRL (treat1*frl: $\beta = 10.21$, $p < .001$). There is also a much smaller positive interaction between two years of treatment and FRL, but it is not significant (treat2*frl: $\beta = 1.16$, $p > .05$). Figure 4.7 below illustrates these positive interaction effects between FRL status and treatment whereby PACE students receiving one year of treatment in the 2015-16 school year are estimated to outperform their non-PACE comparison peers who also qualify for FRL, with all other parameters in the model set to the sample average.

These findings suggest that there is a positive differential effect for PACE students who qualify for FRL, especially those in their first year of treatment. Since this study does not follow students longitudinally (i.e., these are separate cohorts of Grade 8 students), it is unclear whether the positive differential effects exhibited by FRL students after one year of treatment only occur in their first year of exposure to PACE or whether effects accumulate over time.

Figure 4.7 Mean Grade 8 SBAC ELA scale score for free- and reduced-price lunch students by school year and treatment status using the inverse propensity score weighted sample



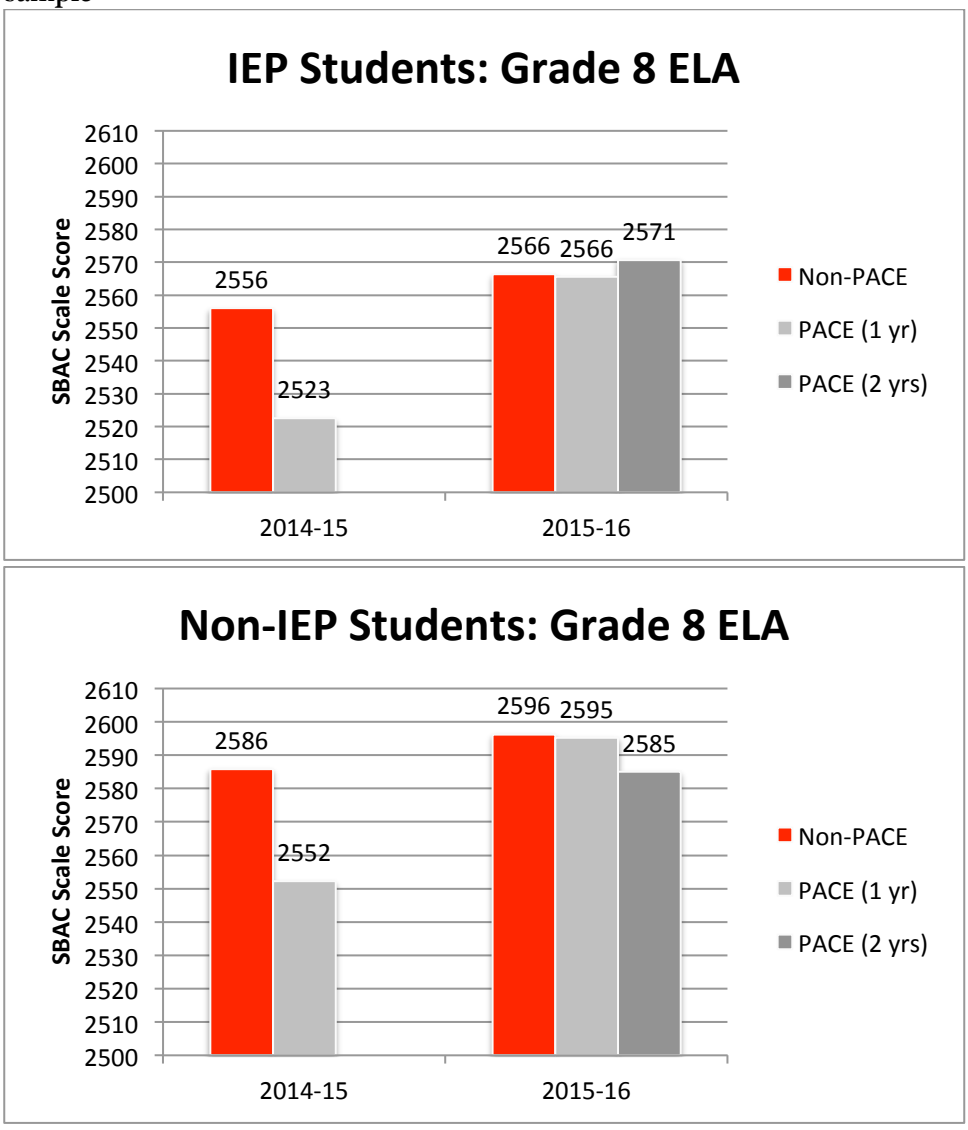
Note. Covariates in the model include student-level characteristics (prior achievement, free-and-reduced price lunch status, disability status, and gender), school-level characteristics (percent of students in the school who are math proficient or above, percent of students in the school who qualify for free-and-reduced price lunch, and number of students in the school who took SBAC), year ID and treatment variables (SBAC Year ID, 1 or 2 years of treatment, interaction between

SBAC Year ID and treatment variables), and cross-level interactions between treatment variables and three student-level characteristics (1 or 2 years of treatment interacted with prior achievement, disability status, and gender).

Disability status. Similar to FRL, the effect of IEP status on Grade 8 ELA achievement varies as a function of treatment. Students who received two years of treatment tend to exhibit a less negative effect of IEP status on ELA achievement, holding all other variables in the model constant (treat2*iep: $\beta = 15.6$, $p < .001$). There is also a very small positive interaction effect between two years of treatment and IEP status, but it is not significant (treat1*iep: $\beta = 0.25$ $p > .05$).

In contrast to the math analyses, when computing effects for the average IEP and non-IEP student by treatment status and year using the parameter estimates from Model 4 in Table 4.8, the achievement gap between IEP and non-IEP students is still evident in the bar graphs (see Figure 4.8). This is because the really large positive interactions between IEP status and treatment for the Grade 8 math weighted sample are not exhibited with the Grade 8 ELA weighted sample.

Figure 4.8 Mean Grade 8 SBAC ELA scale score for IEP students (top panel) and non-IEP students (bottom panel) by school year and treatment status using the inverse propensity score weighted sample

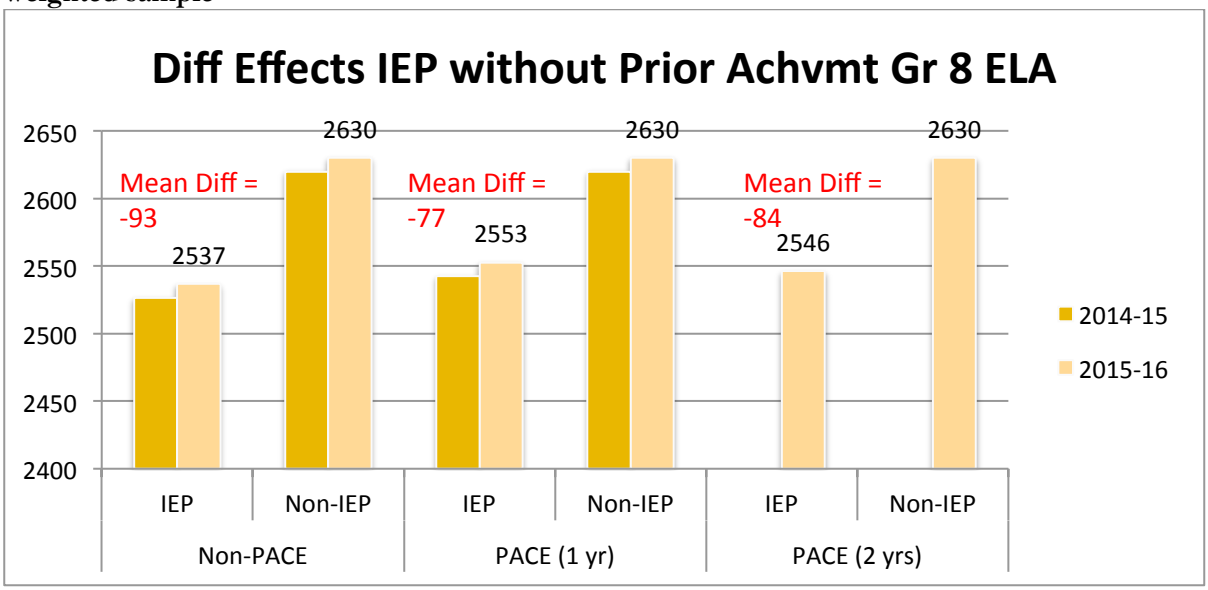


Note. Covariates in the model include student-level characteristics (prior achievement, free-and-reduced price lunch status, disability status, and gender), school-level characteristics (percent of students in the school who are math proficient or above, percent of students in the school who qualify for free-and-reduced price lunch, and number of students in the school who took SBAC), year ID and treatment variables (SBAC Year ID, 1 or 2 years of treatment, interaction between SBAC Year ID and treatment variables), and cross-level interactions between treatment variables and three student-level characteristics (1 or 2 years of treatment interacted with prior achievement, disability status, and gender).

Similar to the Grade 8 math findings, there is also a narrowing of the achievement gap between IEP and non-IEP students for PACE students. Figure 4.9 shows the differential effects of PACE treatment by IEP status, year, and treatment status when prior achievement is not included in

the model[21]. The mean differences between IEP and non-IEP students reduce from around 93 points for the non-PACE comparison group to 77 and 84 points for the PACE groups. These findings suggest a narrowing of the achievement gap by 9 to 16 points for PACE IEP students in comparison to non-PACE IEP students (about 9-16% of the pooled SBAC standard deviation), when controlling for all other student- and school-level characteristics included in Model 4 besides prior achievement. Again, due to the small number of PACE IEP students in the sample, these results should be considered exploratory and in need of replication with a larger sample size and in other grades.

Figure 4.9 Differential effects of IEP status on Grade 8 SBAC ELA achievement using parameter estimates from a model that does not control for prior achievement for the inverse propensity score weighted sample

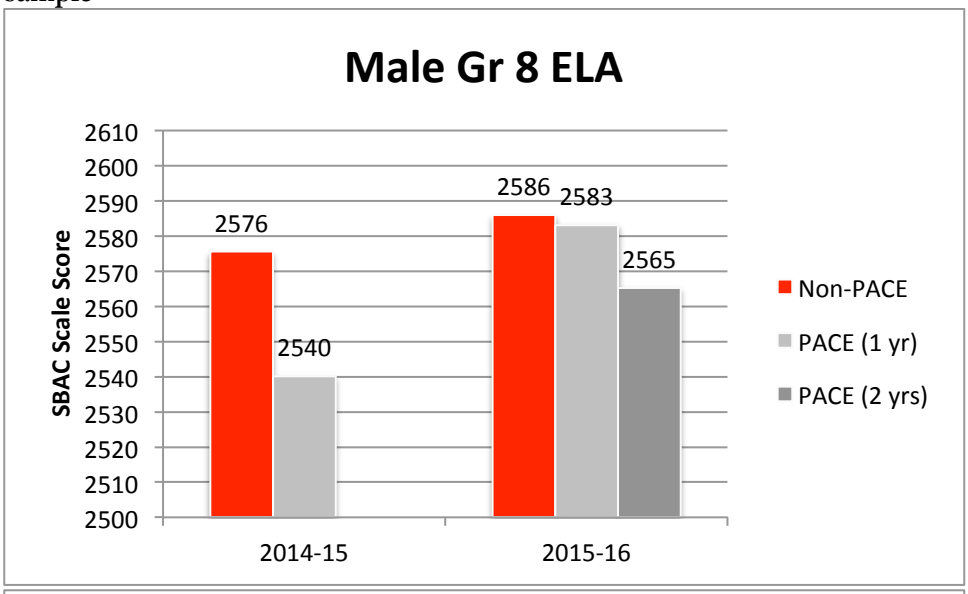


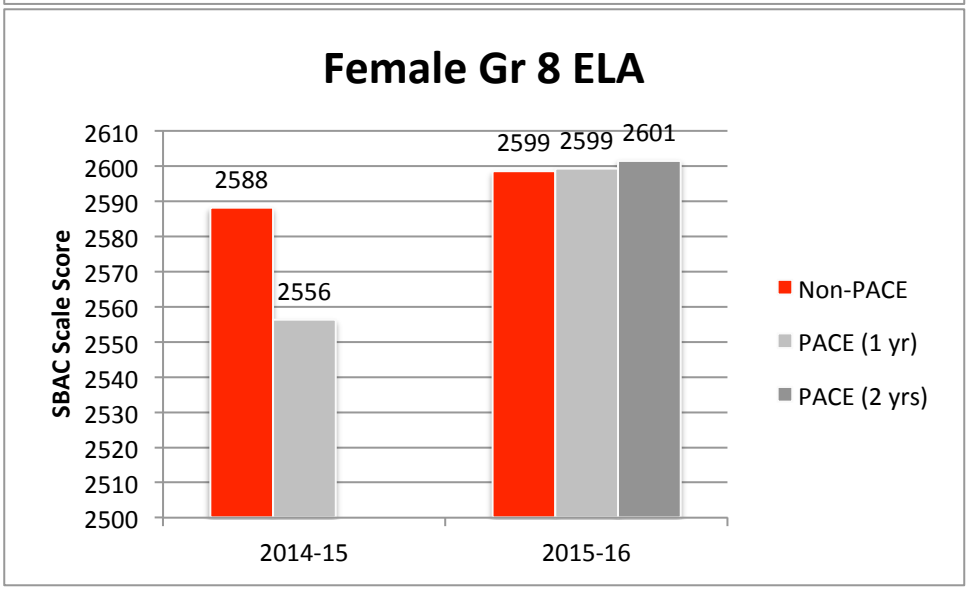
Note. Covariates in the model include student-level characteristics (free-and-reduced price lunch status, disability status, and gender), school-level characteristics (percent of students in the school who are math proficient or above, percent of students in the school who qualify for free-and-reduced price lunch, and number of students in the school who took SBAC), year ID and treatment variables (SBAC Year ID, 1 or 2 years of treatment, interaction between SBAC Year ID and treatment variables), and cross-level interactions between treatment variables and two student-level characteristics (1 or 2 years of treatment interacted with disability status and gender).

[21] Prior achievement was not included in the model in order to isolate the differential effect of treatment for students with disabilities on Grade 8 ELA achievement.

Gender. Also similar to the math findings, female students tend to outperform male students in similar years of the pilot and treatment status. This is because the main effect of gender (male=1; female=0) is negative (gend: β = -12.61, p < .001) as are the interactions between gender and treatment status (treat1*gend: β = -3.57, p > .05; treat2*gend: β = -23.55, p < .001). Figure 4.10 shows the mean Grade 8 SBAC ELA scale score for male students in the top panel and female students in the bottom panel for the average student.

Figure 4.10 Mean Grade 8 SBAC ELA scale score for male students (top panel) and female students (bottom panel) by school year and treatment status using the inverse propensity score weighted sample





Note. Covariates in the model include student-level characteristics (prior achievement, free-and-reduced price lunch status, disability status, and gender), school-level characteristics (percent of students in the school who are math proficient or above, percent of students in the school who qualify for free-and-reduced price lunch, and number of students in the school who took SBAC), year ID and treatment variables (SBAC Year ID, 1 or 2 years of treatment, interaction between SBAC Year ID and treatment variables), and cross-level interactions between treatment variables and three student-level characteristics (1 or 2 years of treatment interacted with prior achievement, disability status, and gender).

As a result of the negative interactions between treatment status and gender, male PACE students are predicted to perform slightly lower than their male non-PACE counterparts in Grade 8 ELA, holding all other variables in the model constant. It is unclear why male students tend to not perform as well on the Grade 8 ELA assessment in comparison to female students. It is also not clear why there is a negative interaction between gender and treatment status. Prior research on performance assessment programs and competency-based education did not examine differences in effects by gender, so it is unclear whether this is a common pattern or not. This could be an area of future research, especially the extent to which this pattern holds over time and in other grade/subject combinations.

Findings for Research Question #3

In order to examine how average treatment effects vary between PACE schools, I used the *lme4* package in R (Bates et al., 2016) to obtain the level-2 residuals by School ID. I am interested in the extent to which PACE schools performed better than predicted (positive residuals) or worse than predicted (negative residuals) in ELA and if there are any patterns across PACE schools or pilot years. In other words, are PACE schools performing better than expected or worse than expected in ELA, based on the level-2 residuals from Model 4? Table 4.9 summarizes the descriptive statistics for the level-2 residuals for Grade 8 ELA.

Table 4.9 Descriptive statistics on level-2 residuals for Grade 8 ELA using the inverse propensity score weighted sample

Mean	0.00
Median	0.79
Std. Deviation	18.49
Range	106.57
Minimum	-65.57
Maximum	41.00

Note. Covariates in the model include student-level characteristics (prior achievement, free-and-reduced price lunch status, disability status, and gender), school-level characteristics (percent of students in the school who are math proficient or above, percent of students in the school who qualify for free-and-reduced price lunch, and number of students in the school who took SBAC), year ID and treatment variables (SBAC Year ID, 1 or 2 years of treatment, interaction between SBAC Year ID and treatment variables), and cross-level interactions between treatment variables and three student-level characteristics (1 or 2 years of treatment interacted with prior achievement, disability status, and gender).

Table 4.10 shows the Grade 8 ELA level-2 residual values for PACE schools by year. Four of the seven PACE schools had positive school-level residuals in Year 2 for ELA. Similar to the Grade 8 math analyses, PACE schools that participated in the pilot for two years tended to exhibit positive residuals in the second year of implementation (e.g., School 20630 and 26505). That said, with only two years of data for three schools it is not possible to make any claims about trends.

Table 4.10 Level-2 residuals for PACE schools by year for Grade 8 ELA using the inverse propensity score weighted sample

School ID	Level-2 Residuals Year 1	Level-2 Residuals Year 2
20270	*	-27.97
20630	-67.88	28.16
20885	*	20.25
22705	-52.37	-66.61
26505	-7.88	10.38
26550	*	-68.21
28400	*	50.55

Note: *=Not yet implementing. Covariates in the model include student-level characteristics (prior achievement, free-and-reduced price lunch status, disability status, and gender), school-level characteristics (percent of students in the school who are math proficient or above, percent of students in the school who qualify for free-and-reduced price lunch, and number of students in the school who took SBAC), year ID and treatment variables (SBAC Year ID, 1 or 2 years of treatment, interaction between SBAC Year ID and treatment variables), and cross-level interactions between treatment variables and three student-level characteristics (1 or 2 years of treatment interacted with prior achievement, disability status, and gender).

Future research could investigate whether these patterns for some schools hold over time and the extent to which they can be explained by contextual within-district implementation factors. There is no apparent relationship between the school-level residuals and fidelity-of-implementation continuum verbalized by the NHDOE (see Table 3.5).

Summary

Overall, findings suggest that PACE students tend to perform lower than their non-PACE comparison peers in Year 1 of the pilot. This most likely reflects that students received only one month of PACE treatment during that school year rather than an implementation dip since the PACE pilot was not officially approved until March 2015—about a month before students took the standardized outcome measure. Findings also suggest that starting in Year 2, there are small positive effects of PACE in Grade 8 math for the average student (d=0.14) for some students, but basically no effect in Grade 8 English language arts.

Results also point to positive differential effects for students with disabilities in Grade 8 math (d=0.20 to 0.50) and Grade 8 ELA (d=0.09 to 0.16), but negative effects for male students that off-set positive treatment effects in Year 2. The findings for students with disabilities should be considered exploratory and in need of replication due to the small sample size in the PACE IEP group. There are mixed and inconclusive findings based on the other student-level characteristics examined—prior achievement and free- and reduced-price lunch.

For schools implementing PACE in both years of the pilot, there is some evidence to suggest that schools perform better than expected starting in the second year of implementation, although the sample size is limited and findings are not generalizable.

In Chapter 5, I discuss the findings across the three research questions in relation to the empirical research literature, explore the implications of these findings for research, policy and

practice, discuss the significance and limitations of this research, and offer recommendations for future research that builds on these findings and research design.

Chapter 5: Discussion, Implications, and Conclusion

Previous chapters in this dissertation provide the background to this study (Chapter 1), situate this study in the context of the previous research literature (Chapter 2), detail the study context and methods (Chapter 3), and present findings related to research questions (Chapter 4). The purpose of this chapter is to discuss the findings in relation to the previous research literature and PACE pilot theory-of-action, discuss the limitations of the study, explore the implications and significance of the findings for research, policy and practice, and offer recommendations for future research that builds on these findings and study design.

Purpose & Overview of the Study

Flat lining or declining achievement over time for K-12 students alongside persistent achievement gaps (Barton & Coley, 2009) prompt policymakers and other education advocates to pursue different paradigms for assessment and accountability in schools. Performance-based assessment, for example, has been promoted for many years as one way to promote deeper learning in schools and also provide useful and timely information to teachers on what students know and can do and at what depth of knowledge (Lane & Stone, 2006; Stecher, 2010). The more recent competency-based education movement dovetails with this call for assessment and accountability reform with a focus on students demonstrating mastery of cognitively complex competencies, flexible pacing and personalized learning that allows students to move on when ready or receive personalized support to re-learn material, and multiple types of assessment (especially performance assessment) to demonstrate proficiency.

And yet for all the interest in performance-based assessment and competency-based education reform, there is little empirical research on these types of reforms. As Shepard and colleagues (1995) state, the benefits of large-scale performance assessment programs have often been inferred from the negative unintended consequences that result from high-stakes testing and

accountability in schools. The same is true for competency-based education—there is very little empirical research on its effectiveness (Freeland, 2014; Haynes et al., 2016; Lewis et al., 2014). The benefits for competency-based education have often been inferred from the negative unintended consequences that result from social promotion policies and inflexible school structures that do not meet each student where they are at and help them to demonstrate proficiency anywhere and anytime. This study addresses these gaps in the knowledge base.

The research literature on performance assessment programs is mainly from the 1990s—a very different policy context—and with little input on the effects that can be expected in the earliest years of the reform. These studies were also designed differently and often focus more on relations between teacher-reported changes in instructional practices and student achievement outcomes (Lane et al., 2002; Parke et al., 2006; Stecher et al., 1998, 2000; Stecher & Chun, 2001; Stone & Lane, 2003). The research literature on recent competency-based education reforms and student outcomes is mainly with charter schools founded as competency-based or personalized learning schools (Bill & Melinda Gates Foundation, 2014; Pane et al., 2015)—although there is one study with public schools, but the intervention dosage was very minimal and involves only a handful of teachers (Steele et al., 2014). This limits what can be inferred for public schools and public school students, as well as interventions with more implementation fidelity and scope. This study builds upon the prior literature as it anticipates these design challenges and limitations. This study also extends the prior literature in examining differential policy effects for certain subgroups of students.

The policy context of this study is recent federal legislation, the *Every Student Succeeds Act* of 2015, that allows up to 7 states to pilot innovative assessment and accountability systems. Innovative systems may address lackluster student achievement gains and equity concerns while promoting novel solutions to what many consider to be this nation's over-reliance on high-stakes achievement testing in K-12 schools. Before the innovative pilot launches, however, selected New Hampshire

school districts are involved in a proof of concept model whereby determinations of student proficiency in grades 3-8 and once in high school in math and English language arts are made using a combination of local, common, and state-level assessments.

New Hampshire's Performance Assessment of Competency Education (PACE) pilot was officially approved by the USDOE in March 2015 and, as of the writing of this dissertation, is now in its fourth year of implementation (2014-15, 2015-16, 2016-17, and 2017-18 school years). This study examines the first two years. There is a state-level achievement test administered once per grade span in the NH PACE pilot that acts as an external audit on the system. In every other grade and subject combination, school districts use local assessments (including performance assessments) and one common performance assessment alongside teacher judgment surveys to determine student proficiency. This innovative assessment and accountability system incorporates both performance-based assessment and competency-based education.

Although the desire to meaningfully prepare all students for college or career is at the heart of NH's PACE pilot, there is no empirical evidence to date on the extent to which the PACE pilot is improving student achievement outcomes as measured by a state level achievement test—a proxy for college and career readiness. There is also no empirical evidence on how specific subgroups of students such as students with disabilities or students who qualify for free- and reduced-price lunch are impacted by such a reform. And yet as policymakers and educators across the nation explore innovative approaches to assessment and accountability they want to know how these systems may impact teaching and learning.

The purpose of this study, therefore, was to examine the effect of the PACE pilot on student achievement outcomes. There were three research questions. First, what is the average effect of the PACE pilot on Grade 8 student achievement in mathematics and English language arts when comparing students with similar probabilities of being selected into the PACE pilot? Second, do

those average effects vary for certain subgroups of students: students with disabilities, males/females, free- and reduced-price lunch students, or students with low or high prior achievement? And, finally, how do average effects vary among PACE schools? Are there some PACE schools who perform better than expected while other schools are performing worse than expected and is there any pattern to those differences?

In order to examine these three research questions, it was first important to establish equivalent treatment and comparison groups at baseline in order to address the likely selection bias inherent in the PACE group. Districts and schools self-selected into the PACE pilot and there are pre-existing differences between PACE and non-PACE districts and schools that are likely related to both selection and student outcomes. These pre-existing differences potentially bias effect estimates and threaten the internal validity of the study. Therefore, inverse propensity score weighting was used to create roughly equivalent groups at baseline based on observable district-level characteristics of the students in the PACE group and non-PACE group. Since students are nested within districts/schools, multi-level modeling was then used with the inverse propensity score weights to examine the effects of one or two years of treatment on Grade 8 student achievement outcomes in math and English language arts. Interactions between treatment and student-level characteristics were also examined to investigate whether effects varied for different subgroups of students. Random intercept effect estimates were used to examine whether PACE schools performed better or worse than predicted and if there were any patterns in those school-level residuals.

It is important to note that because baseline equivalence standards were not met between the PACE and non-PACE group, this study is purely descriptive and observational. For example, the PACE group had higher percentages of students with disabilities and free- and reduced-price lunch, as well as lower percentages of student who were proficient or above in math and English language arts. Effect estimates, therefore, likely underestimate treatment rather than overestimate.

Summary of Findings

Research Question #1

Overall, findings suggest that PACE students tend to perform lower than their non-PACE comparison peers in Year 1 of the pilot. This most likely reflects that students received only one month of PACE treatment during that school year rather than an implementation dip since the PACE pilot was not officially approved until March 2015. Findings also suggest that starting in Year 2, there are small positive effects of PACE in Grade 8 math for the average student with one year of treatment (d=0.14), but basically no effect in Grade 8 English language arts.

Research Question #2

There were two subgroups of students that exhibited differential effects in both subject areas: students with disabilities and male students. First, findings suggest that there are positive differential effects for PACE students with disabilities in both subject areas. For example, there is evidence to suggest that students with disabilities tend to exhibit a positive differential effect of PACE treatment in both Grade 8 math (20-50% of a standard deviation) and Grade 8 ELA (9-16% of a standard deviation). These positive differential effects significantly narrow the achievement gap between IEP and non-IEP students for PACE students with disabilities in comparison to non-PACE students with disabilities; however, caution should be taken in extrapolating from these findings as results are based on a small number of students.

Second, findings suggest that male students tend to exhibit negative differential effects of treatment. For example, male students receiving PACE treatment are estimated to perform about the same in Grade 8 math and slightly lower in Grade 8 ELA as their non-PACE, male comparison peers starting in Year 2, on average. This is because the positive average effects of PACE starting in Year 2 that were noted above are off-set for male students because of the negative interactions.

There were inconclusive and mixed findings for differential effects based on student socioeconomic status and prior achievement. This study found that only Grade 8 ELA students who qualify for free- and reduced-price lunch tended to exhibit a positive differential effect of PACE treatment around 6% of a standard deviation—but not in Grade 8 math. There were also mixed and inconclusive findings related to differential effects for students based on prior achievement. There were very small differential effects of prior achievement for some PACE students that benefited lower performing students who received one year of treatment, but higher performing students who received two years of treatment. These findings make it difficult to make any generalizations based on these two student-level characteristics.

Research Question #3

For schools implementing PACE in both years of the pilot, there is some evidence to suggest that schools perform better than expected starting in the second year of implementation. For example, two of the three PACE schools that implemented for both years of the pilot had mean Grade 8 SBAC school-level achievement that was lower than predicted in the first year of implementation, but better than predicted in the second year of implementation. Also, there is some evidence to suggest that more PACE schools perform better than expected in Grade 8 ELA in comparison to Grade 8 math; however, it is impossible to make any generalizations from this data because there is only two years of data for three schools and one year of data for four schools.

Discussion of Findings

Average Effects in Grade 8 Math and English Language Arts

The findings that there were small positive effects for some PACE students in Grade 8 math (d=0.14) and no effect in Grade 8 ELA after two years of implementation, suggest that PACE students are not "harmed" academically as part of their school's involvement in the PACE pilot. An

inference could be made from these effects that PACE students are provided an equitable opportunity to learn the content standards.

However, one of the difficulties encountered in this type of observational research is that there are other reforms taking place in PACE schools and districts (and in NH generally) that may be interacting with PACE effects. For example, two PACE districts (111 and 476) received K-8 math professional development during the 2015-16 school year through the Ongoing Assessment Project (OGAP Math, n.d.). OGAP trains teachers to use formative assessments and analyze student thinking relative to mathematical learning progressions in order to guide instructional steps. It is impossible to disentangle the effects of these other school- or district-level reforms from the effects of the PACE pilot on student achievement outcomes.

That said, these findings mirror earlier research on classroom performance assessments and competency-based education. For example, Shepard and colleagues (1995) found similar gains after one year in mathematics (d=0.13) and no effect after one year in English language arts. Pane and colleagues (2015) found effects stronger in magnitude in math (d=~0.20) than reading (d=~0.14) in the first three years of the competency-based education reform. These findings also mirror earlier research on Maryland's statewide performance assessment program used for accountability purposes in the 1990s (Lane et al., 2002; Parke et al., 2006; Stone & Lane, 2003). In studies on that program, the general trend was a significant increase in mean school-level performance over the 5-year time period except in writing (Stone & Lane, 2003). There was a slight dip in writing in the early years followed by a steady increase over the last 3 years of the study. This suggests that it may be more common to see positive effects of a pilot with a strong focus on performance assessments such as the PACE pilot after one year in Grade 8 math, but it may take longer for effects to accumulate in English language arts.

This dissertation raises the question about why effects in Grade 8 math may be exhibited before effects in Grade 8 English language arts in these types of reforms—assuming effects are due to PACE alone and not to other contextual factors or reforms taking place in these schools/districts. One reason why positive effects may appear in math earlier than in English language arts is that instructional practices in math may be more affected by the PACE theory-of-action. For example, math instruction can focus on lower depth of student understanding such as basic recall and computing procedures with less attention to the deeper mathematical concepts and skills. The PACE theory-of-action postulates that performance-based assessment within a competency-based learning environment impacts the instructional core of classroom practices such that teachers focus on higher-order math thinking skills such as application, synthesis, evaluation, and analysis rather than having students just memorize basic math facts and recall mathematical procedures. It may be the case that those types of reform-oriented instructional changes are more substantial in math than in reading and represent a greater divergence from prior instructional strategies. If it is the case that there are greater impacts of the PACE pilot on math instruction in the classroom then it makes sense that effects would likely appear earlier in the reform in math. If the results in Maryland are any indication (Stone & Lane, 2003), it is likely that there will be positive effects in English language arts over time for students in the PACE pilot, but it may take longer for effects to accrue over time in English language arts as reform-oriented instructional changes may not be as drastic as in math. Research using other grade levels will help contextualize these findings and provide insight into how they generalize beyond Grade 8.

Since it is not uncommon to see little impact on student achievement from major reform efforts during the first 3-5 years of implementation (Fullan, 2001), the fact that a positive effect was exhibited by some PACE students in Grade 8 math after only two years of implementation may bolster support for the underlying PACE theory-of-action. Furthermore, given evidence from the

organizational change and management literature that it is not uncommon for performance dips to occur for a short period after major organizational changes (Herold & Fedor, 2008; Jellison, 2006), it is likely that basically no implementation dip outside of Year 1 where it could be argued there was very little treatment would reflect positively on the PACE pilot.

In general, these findings add to the research base about how performance assessment programs and competency-based education reforms affect student performance in the first few years of implementation—a gap in the prior research literature. These findings also provide information about differences in average effects by subject area and how effects in math may show up earlier than in English language arts in similar reforms. This is an area where future research could examine other grade levels to see if these effects generalize across grades. Future research could also collect local classroom assessment artifacts from math and English language arts in order to examine the alignment between the assessments and the depth and breadth of the content competencies. Due to the rolling cohort nature of PACE implementation, it may also be possible to examine differences in the quality of local classroom assessments at the school- or district-level based upon length of time in the PACE pilot—and even how instructional practices change from prior to implementation (Tier 2 and 3) to implementation (Tier 1).

Differential Effects for Certain Student Subgroups

The findings related to differential effects by subgroup are exploratory as there are small sample sizes for some subgroups. That said, findings suggest that students with disabilities attending PACE schools exhibit achievement gains 20-50% of a standard deviation in Grade 8 math and 9-16% of a standard deviation in Grade 8 English language arts. These achievement gains for students with disabilities significantly narrow the achievement gap between non-IEP and IEP students for PACE students. This is an important area of future research. For example, do these findings hold across grade levels and when using analyses with larger sample sizes?

Conceptually these findings fit well with the PACE theory-of-action because the use of competency-based approaches to curriculum and instruction alongside the use of curriculum-embedded, high quality performance based assessments is intended to drive changes to the instructional core of classroom practices such that student achievement improves. For students with disabilities, this may mean that curriculum and instruction is differentiated to meet them where they are in their development of grade and subject area competencies and then offers them multiple pathways to demonstrate proficiency. Students with disabilities who are not able to demonstrate competency also have access to timely, differentiated, and individualized support mechanisms that target their misunderstandings. These support structures may benefit students with disabilities as they are provided other opportunities to master key competencies, multiple opportunities to demonstrate mastery, and multiple types of assessments to show their learning.

The use of performance-based assessment may also benefit students with disabilities because it affects both the process and the product of assessing student progress towards proficiency. Students are provided with qualitative descriptions of performance from beginning levels of understanding to advanced levels of understanding on multi-dimensional, analytic rubrics. These same rubrics that are provided in advance are also used to provide specific, meaningful, and relevant feedback to students on what they know and can do and at what depth of knowledge. Teachers are also aided in their process of instruction because the performance assessment itself provides specific information on student misunderstandings and target areas for re-teaching that is personalized to the student.

In all these ways, it conceptually makes sense that students with disabilities tend to exhibit positive differential effects from participating in PACE because of specific, timely, and relevant feedback on performance provided to students as well as congruence among the curriculum, instruction, and assessment components. However, it is difficult to contextualize these study

findings because interactions between disability status and treatment were not tested in the prior research literature on the efficacy of performance assessment programs or competency-based education reforms. Therefore, it is unclear whether this pattern is consistent with similar reform efforts or whether this finding differs.

Future research could examine other grade levels and explore differential effects with other methods. Future research could also take a deep dive in to PACE schools/districts and qualitatively explore the processes and procedures special education teachers, general classroom teachers, and paraprofessionals are using with the IEP students and the extent to which those processes and procedures may help explain these differential effects. Disaggregating student scores according to their special education category may also yield useful information regarding subpopulations of students and differential effects of assessment and accountability policies/programs.

The only differential effects noted in the prior literature were for lower achieving students and socioeconomically disadvantaged students who appeared to benefit more from treatment than their high achieving and economically advantaged peers—although the prior literature did not actually use interaction terms in multivariate analyses to examine differential effects (e.g., Bill & Melinda Gates, 2014; Pane et al., 2015).

In line with those findings from the prior literature, this study found that students who qualify for free- and reduced-price lunch tended to exhibit a positive differential effect of PACE treatment for Grade 8 English language arts students around 6% of a standard deviation. However, different from the prior literature, there were mixed findings related to differential effects for students based on prior achievement. There were very small differential effects of prior achievement for some PACE students that benefited lower performing students who received one year of treatment, but higher performing students who received two years of treatment. It may be that

results from this study varied from the prior literature because interaction terms and hypothesis testing was employed in this study whereas, in the early studies, mainly descriptive results were used.

The final significant interaction effect between a student-level characteristic and treatment status was for male students. Grade 8 male students tended to exhibit negative differential effects of treatment in both math and English language arts. It is unclear why there is a negative interaction between gender and treatment status because differential effects by gender were not tested in the prior research literature reviewed in this paper. This makes it unclear whether this effect somehow pertains to this population and not others, or if these negative effects for male students are more widespread. Nationally, males tend to score about the same as female students in Grade 8 math (National Center for Education Statistics, 2017), but a little lower in Grade 8 ELA, on average (National Center for Education Statistics, 2015). Further research on reform systems that are designed using performance-based assessments and/or competency-based education could explore these relationships with different populations, subject areas, and grade levels.

Differences in School-Level Effects Among PACE Schools

In terms of differences in school-level effects among PACE schools, there is some evidence to support the claim that a school performs better than predicted the longer it implements the PACE pilot. As stated previously, it is difficult to generalize from this data because there is only two years of data for three schools and one year of data for four schools. Examining school-level achievement trends over time and in other grades will help elucidate whether there are any patterns in performance based on years of implementation, grade, and/or subject area. Furthermore, follow-up research could investigate whether "highly successful" PACE schools employ qualitatively different curricular, instructional, or assessment practices than "not as successful" PACE schools.

Limitations

It is worth noting that the New Hampshire context—though important in its own right—may not be representative of other states nationwide and therefore treatment effects may differ in different contexts. Along these lines, because of the low percentages of racial/ethnic minorities and limited English proficient students in the state of New Hampshire, this study cannot illuminate the effect the PACE pilot has on these diverse student groups and achievement gaps. Future research could be conducted in other states and settings with a more ethically/racially diverse student population to examine effects in those contexts.

Additionally, one challenge in terms of extrapolating from these findings any conclusions about performance assessment programs or competency-based education policy reforms is that it is impossible to disentangle the effects of each reform, or other reforms taking place simultaneously within districts. This means it is impossible to isolate the effects of the PACE pilot on student achievement outcomes and the theoretical aspects undergirding the PACE pilot theory of action are confounded in this study and cannot be analyzed separately.

Also, as in any new educational program/policy, there are differences in organizational capacity, leadership, and implementation that affect program/policy outcomes. For example, the fidelity-of-implementation among the PACE districts is unknown at this time and most likely varies district-to-district and even between schools within districts. It is possible that effects vary as a function of how the PACE pilot is implemented in a district and/or school, which is not accounted for in this study. We know from the research literature that fidelity of implementation is an important factor that can explain why a program in one location is considered effective, while the same program in another location is not effective (Fullan, 2001). Also, according to implementation science research, how a program is remade and adapted in local contexts can also explain variability in program effects (Durlak & DuPre, 2008); however, none of this is accounted for in this study.

Future research could conduct research on the levels of implementation along the key dimensions of the PACE theory of action and use some type of implementation metric as a variable to explain differences in student achievement outcomes.

Another limitation of this study is the small PACE sample size. As noted in Table 3.11, there were only three PACE schools and 456 students who received two years of treatment and seven PACE schools and 681 students who received one year of treatment. The subgroup analyses spliced these groups into even smaller numbers—particularly the analyses of students with disabilities and the analyses of free-and-reduced lunch students. More research needs to be done with larger sample sizes over time, in other grades, and with different analyses to verify the findings in this dissertation.

One other important limitation to this study is that neither students nor schools were randomly assigned to their treatment status. Districts volunteered to be part of the PACE pilot, and only students who lived in those school districts were part of the pilot intervention. While I used propensity score methods in an attempt to address this selection bias, attempts to create equivalent treatment and comparison groups at baseline based on observable district-level covariates plausibly related to both selection and outcomes were not totally successful. It is therefore possible that students and schools may have differed in unobserved ways that were related to both their selection (involvement in the PACE pilot) and the measured student achievement outcomes of interest. As a result, this study cannot make any causal claims. It is critical to remember that this study provides a descriptive (non-causal) examination of student outcomes following exposure to the first two years of the PACE pilot program.

Implications for Research

As mentioned previously, there has been a lot of interest over time in utilizing large-scale performance assessment programs and competency-based models of education to close achievement gaps, prepare students for college or career, and facilitate teaching and learning of more cognitively

complex competencies. And yet, as elucidated in Chapter Two, there is not a lot of empirical evidence on the efficacy of these reforms to improving student achievement outcomes. This study begins to fill the gap in the research base on the efficacy of performance assessment programs and competency-based education to improving student achievement outcomes.

However, there is still a lot left to understand about the effects of an innovative assessment and accountability system such as NH's PACE pilot on student achievement outcomes, as well as many other outcomes. For example, this study focused on only Grade 8 mathematics and English language arts student achievement; future research could explore other grades and subject areas. Also, because the PACE pilot continues to scale up to include other schools/districts each year and there will be a new state level achievement test in NH that replaces SBAC starting in spring 2018, future research could explore treatment effects using a standardized outcome measure and different methods. There are also many other outcomes of interest that could be examined such as student motivation and engagement, long-term postsecondary outcomes such as going to college, staying in college, and graduating from college, and even rates of remedial college coursework for PACE students (to name a few). Furthermore, New Hampshire is a unique context as it has low percentages of minority students and students with limited English proficiency. New Hampshire is also a state with a relatively high average median household income and large rural/small town population with only a couple semi-urban areas. There is a need for continued research in other contexts with more diversity in order to examine how these types of reforms may impact different settings.

It is important that future research drills down to the classroom level and examine alignment between local summative assessments given within PACE schools and the breadth and depth of state model competencies or content standards. This would provide evidence about how the PACE theory-of-action is impacting (or not impacting) the instructional core of classroom practices such

that teachers are teaching and assessing students at a greater depth of understanding. This future research could also include examining an alignment index in relation to student outcomes.

Due to the positive differential effects noted for students with disabilities, follow-up research with other grades and larger sample sizes is needed to support these findings. Future research could also explore how effects vary for students in different disability categories. Students with disabilities across the state could also be surveyed about how their teachers meet their learning needs and about their levels of engagement and motivation in school. Differences in survey responses between PACE and non-PACE students with disabilities could be examined to see if there are any significant differences in students' perceptions. Also, qualitative research could be conducted in PACE schools/districts to explore how special education teachers and other professionals who work directly with students with disabilities modify their curricular and instructional strategies to meet diverse learning needs. For example, are there high leverage practices that PACE schools/districts are employing with their students with disabilities that may be leading to these positive differential effects and how do those practices relate to policies around performance-based assessment and competency-based education? Students with disabilities could also be interviewed about their perceptions and how their school/district has adapted and adjusted their program since joining the PACE pilot.

There is also a need for future research on the negative differential effects noted for male students. It is unclear from this research why male students attending PACE schools tend to exhibit lower performance, on average, than their male counterparts attending non-PACE schools. Similar to the recommended research for investigating differential effects for students with disabilities, this would be an area where survey research and qualitative research may elucidate why these achievement patterns are exhibited in this study.

Future research could also examine student and school performance growth over time. For example, there is the potential for a cumulative effect of instruction in the PACE environment on student achievement. Students in this study were exposed to this type of instruction in the middle of their education. It may be the case that students who begin their education in this type of environment would show increasing impact over time. Future research could investigate this potential cumulative effect.

Finally, future research could examine how student and school-level effects vary based on fidelity-of-implementation and reform-oriented changes to instructional practice. For example, there has been little empirical research on the implementation of competency-based education models, although research is starting to emerge (Haynes et al., 2016; S. Ryan & Cox, 2017; Steele et al., 2014) and more calls for research remain (Freeland, 2014; Le et al., 2014). Because large-scale performance assessment programs were discontinued once *No Child Left Behind* was implemented, there has also been no empirical research on the implementation of those programs in the last fifteen years and a recent formative evaluation on the NH PACE pilot was all I could locate that examined implementation (Becker et al., 2017).

Implications for Policy

At the end of the day, one thing policymakers and other stakeholders want to know is—does this innovative assessment and accountability policy have the intended effect, and for whom? Is this policy leading to harm or leading to benefit? Specifically, are students in the pilot provided an equitable opportunity to learn the content standards? These are the policy outcomes or the intended and unintended consequences of a policy for those on the receiving ends.

Since the NH PACE pilot operates under a waiver from federal statutory requirements related to state annual achievement testing, part of the conditions for continuing the waiver stipulated by the U.S. Department of Education is that the NHDOE demonstrate that all students

who participate in the pilot are provided and equitable opportunity to learn based on the criterion of "no harm" on the state achievement test. Findings that many PACE students performed just as well or slightly better as students in the comparison group on standardized measures starting in the first full year of implementation alongside the fact that there was no evidence for an implementation dip, makes a strong case that the PACE pilot has met the criterion of "no harm" on the state achievement test. Key stakeholders and policymakers could use the findings from this study to support the claim that students who attend districts or schools involved in the innovative assessment and accountability pilot are provided an equitable opportunity to learn the content standards.

The PACE pilot is closely watched by educators and policymakers nationwide as a potential model of what an innovative assessment and accountability system might look like, particularly one that utilizes performance assessment within a competency-based education framework (Rothman & Marion, 2016). The PACE pilot addresses national concerns about over-testing, the negative effects of high-stakes testing and accountability on teaching and learning, and the need for systemic educational change to close achievement gaps. Results from this study may provide the empirical evidence and political capital others states need to move forward with their own plans to design, apply for, and implement an innovative pilot and/or enact legislation that promotes competency-based education. For example, currently there are only 10 states with "comprehensive policy alignment" with competency-based models of education (Sturgis, 2016). This means that there are only a handful of states whose policies specifically require that students graduate not based on credit hours, but based on demonstration of proficiency related to state content standards or competencies. This research may inform future state-level policies related to competency-based education and the importance of students demonstrating mastery, proficiency, or competency rather than sitting in a seat for a certain number of days or hours per year.

This research may also inform the use of top-down accountability mandates as a policy lever to effectuate systemic school reform. For example, the PACE theory-of-action focuses on reciprocal accountability rather than external rewards and sanctions to accomplish organizational change and growth. Districts and schools are provided capacity building supports and resources from the state to implement the innovative system and tasked with the responsibility of holding themselves accountability for student growth. The PACE system promotes a very different accountability model than the *No Child Left Behind Act* where there were specific sanctions faced by schools that did not meet adequate yearly progress and continues to a lesser extent under the *Every Student Succeeds Act*.

Implications for Practice

In terms of implications for practice, this observational study provides initial empirical evidence that learning gains exhibited by students resulting from these types of reforms may be transferring or carrying over to a very different assessment of student proficiency—the state achievement test. This transfer of subject matter knowledge and skills in one context to another context is exactly what reformers envision because transfer signals that deeper learning has taken place. In other words, knowledge and skills taught in one setting can be applied in another setting equally well, especially on a state achievement test that is designed to measure the breadth and depth of the content standards. It will be important to examine student achievement trends over time and in other grades to investigate whether these early effects continue over time and are exhibited in other grades.

The fact that many students perform equally well or slightly better on the state achievement test also implies that content coverage in PACE schools is not sacrificed for the sake of content depth. Finding the balance between content coverage and content depth was a concern noted in the prior research literature on mastery learning and why there was a debate over the type of outcome measure that should be used to measure program effects. Similarly, because competency-based

education approaches require students to demonstrate proficiency in order to move on in the curriculum, there are equity concerns about how that type of learning model may affect certain subgroups of students who may struggle to demonstrate proficiency. There is no evidence from findings in this study that students who may struggle to demonstrate proficiency (such as students with disabilities) are negatively impacted by participation.

Conclusion

Many schools, districts, and states across the United States pursue reforms and implement policy changes around competency-based education and performance-based assessments because of the belief that doing so will improve overall student achievement, narrow or close achievement gaps and help all students to succeed in college or career. In other words, excellence and equity concerns drive many of the policy decisions that lead to similar reforms as those implemented in New Hampshire's PACE pilot.

The significance and contribution of this study to the research literature is that it answers a primary question policymakers and other stakeholders want to know early in the implementation of any major reform initiative—is there any evidence that the policy is having its intended effect? Findings from this study are not conclusive, but do provide modest evidence that the PACE pilot is having a positive effect on Grade 8 student achievement outcomes in mathematics starting in the second year of implementation and no effect on Grade 8 English language arts outcomes. Findings could be used to provide assurance to key stakeholders that PACE students are provided an equitable opportunity to learn the content standards.

REFERENCES

Allensworth, E. M., Moore, P. T., Sartain, L., & Torre, M. (2016). The educational benefits of attending higher performing schools: Evidence from Chicago high schools. *Educational Evaluation and Policy Analysis*. http://doi.org/10.3102/0162373716672039

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Education Research Association and National Academy of Education.

Anderson, L. W., & Burns, R. B. (1987). Values, evidence, and mastery Learning. *Review of Educational Research*, *57*(2), 215–223. http://doi.org/10.3102/00346543057002215

Anderson, S. A. (1994). Synthesis of research on mastery learning. Northville, MI: Northville Public Schools.

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher*, *36*(5), 258–267. http://doi.org/10.3102/0013189X07306523

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424. http://doi.org/10.1080/00273171.2011.568786

Baker, E. L., & Gordon, E. W. (2014). From the assessment OF education to the assessment FOR education: Policy and futures. *Teachers College Record*, *116*(11).

Barton, P. E., & Coley, R. J. (2009). Parsing the achievement gap II. Princeton, NJ: Educational Testing Service.

Bates, D., Maechler, M., Bolker, B., Walker, S., Bojesen, R. H., Singmann, H., … Green, P. (2016). Linear mixed-effects models using "Eigen" and S4. Retrieved from http://lme4.r-forge.r-project.org/

Becker, D. E., Thacker, A. A., Sinclair, A., Dickinson, E. R., Woods, A., & Wiley, C. R. H. (2017). Formative evaluation of New Hampshire's Performance Assessment of Competency Education (PACE), Final Report. Alexandria, VA: HumRRO, Center for Innovation in Education.

Bennet, R. E. (2014). Preparing for the future: What educational assessment must do. *Teachers College Record*, *116*(11).

Bill & Melinda Gates Foundation. (2014). Early progress: Interim research on personalized learning. Seattle, WA: RAND Corporation and Bill & Melinda Gates Foundation.

Block, J. H. (1978). The " C " in CBE. *Educational Researcher*, 13–16.

Block, J. H., & Anderson, L. W. (1975). *Mastery learning in classroom instruction*. New York, NY: Macmillan.

Block, J. H., & Burns, R. B. (1976). Mastery learning. *Review of Research in Education*, *4*, 3–49.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment*, *1*(2).

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal*, *42*(2), 231–268. http://doi.org/10.3102/00028312042002231

Borko, H., & Elliott, R. (1998). Tensions between competing pedagogical and accountability commitments for exemplary teachers of mathematics in Kentucky (CSE Technical Report 495). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Borko, H., Elliott, R., & Uchiyama, K. (2002). Professional development: A key to Kentucky's educational reform effort. *Teaching and Teacher Education*, *18*, 969–987. http://doi.org/10.1016/S0742-051X(02)00054-9

Bramante, F., & Colby, R. L. (2012). *Off the clock: Moving education from time to competency*. Thousand Oaks, CA: Sage.

Brown, R. S., & Coughlin, E. (2007). The predictive validity of selected benchmark assessments used in the Mid-Atlantic region (Issues & Answers Report, REL 2007-No.017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Educaiton Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

Chattergoon, R., & Marion, S. F. (2016). Not as easy as it sounds: Designing a balanced assessment system. *National Association of State Boards of Education*, *16*(1), 6–9.

Chatterji, M. (2002). Models and methods for examining standards-based reforms and accountability initiatives: Have the tools of inquiry answered pressing questions on improving schools? *Review of Educational Research*, *72*(3), 345–386. http://doi.org/10.3102/00346543072003345

City, E. A., Elmore, R. F., Fiarman, S. E., & Teitel, L. (2009). *Instructional rounds in education: A network approach to improving teaching and learning*. Cambridge, MA: Harvard Education Press.

CompetencyWorks. (2011). What is competency education? Washington, DC: CompetencyWorks and iNACOL.

CompetencyWorks. (2014). Understanding competency education in K-12: What is competency education? Retrieved April 25, 2015, from http://www.competencyworks.org/wp-content/uploads/2014/09/CWorks-Aligning-Federal-Policy.pdf

Cotton, K., & Savard, W. G. (1982). Mastery learning: Topic summary report. Research on school effectiveness report. Portland, OR: Northwest Regional Educational Laboratory.

Council of Chief State School Officers. (2015). Comprehensive statewide assessment systems: A framework for the role of state education agency in improving quality and reducing burden. Washington, DC: Author.

Cronin, J., Kingsbury, G. G., McCall, M. S., & Bowe, B. (2005). The impact of the No Child Left Behind Act on student achievement and growth: 2005 edition. Portland, OR: Northwest Evaluation Association.

Darling-Hammond, L., Wilhoit, G., & Pittenger, L. (2014). Accountability for college and career readiness: Developing a new paradigm. *Education Policy Analysis Archives*, *22*(86). Retrieved from http://dx.doi.org/10.14507/epaa.v22n86.2014

Davey, T., Ferrara, S., Holland, P. W., Shavelson, R., Webb, N. M., & Wise, L. L. (2015). Psychometric considerations for the next generation of performance assessment. Princeton, NJ: Educational Testing Service.

Diamond, J. B., & Spillane, J. P. (2004). High-stakes accountability in urban elementary schools: Challenging or reproducing inequality? *Teachers College Record*, *106*(6), 1145–1176. http://doi.org/10.1111/j.1467-9620.2004.00375.x

Doorey, N., & Polikoff, M. (2016). Evaluating the content and quality of next generation assessments. Washington, DC: Thomas B. Fordham Institute.

Downs, A. (1972). Up and down with ecology--the "issue-attention cycle." *Public Interest*, *28*(Summer), 38–50.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*(3–4), 327–350. http://doi.org/10.1007/s10464-008-9165-0

Earl, L. M. (2003). *Assessment as learning: Using classroom assessment to maximize student learning* (1st ed.). Thousand Oaks, CA: Corwin Press.

Elmore, R. F. (2004). Moving forward: Refining accountability systems. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 276–296). New York, NY: Teachers College Press.

Evans, C. M., & Lyons, S. (2017). Comparability in balanced assessment systems for state accountability. *Educational Measurement: Issues and Practice*. http://doi.org/http://dx.doi.org/10.1111/emip.12152
Every Student Succeeds Act. (2015). Pub.L. 114-95 § 114 Stat. 1177.
Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-Based Assessment and Instructional Change: The Effects of Testing in Maine and Maryland. *Educational Evaluation and Policy Analysis*, *20*(2), 95–113. http://doi.org/10.3102/01623737020002095
Freeland, J. (2014). From policy to practice: How competency-based education is evolving in New Hampshire. Lexington, MA: Clayton Christensen Institute for Disruptive Innovation. Retrieved from www.christenseninstitute.org
Fuhrman, S. H., & Elmore, R. F. (Eds.). (2004). *Redesigning accountability systems for education*. New York, NY: Teachers College Press.
Fullan, M. (2001). *Leading in a culture of change*. San Franscisco, CA: Jossey-Bass.
Funk, M. J., Westreich, D., Wiesen, C., Sturmer, T., Brookhart, M. A., & Davidian, M. (2011). Double robust estimation of causal effects. *American Journal of Epidemiology*, *173*(7), 761–767. http://doi.org/10.1093/aje/kwq439
Gearhart, M., & Herman, J. L. (1995). Portfolio assessment: Whose work is it? Issues in the use of classroom assignments for accountability. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.
Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education*, *24*(1), 355–392. http://doi.org/10.3102/0091732X024001355
Gong, B. (2010). Using balanced assessment systems to improve student learning and school capacity: An introduction. Washington, DC: Council of Chief State School Officers and Rennaissance Learning.
Graham, S. E., & Kurlaender, M. (2011). Using propensity scores in educational research: General principles and practical applications. *The Journal of Educational Research*, *104*(5), 340–353.
Guo, S., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications* (2nd editio). Thousand Oaks, CA: Sage.
Guskey, T. R. (1986). Defining the critical elements of a mastery learning program. In *Paper presented at the Annual Meeting of the American Educational Research Association*. San Franscisco, CA.
Guskey, T. R. (1987). Rethinking mastery learning reconsidered. *Review of Educational Research*, *57*(2), 225–229. http://doi.org/10.3102/00346543057002175
Guskey, T. R. (1994). Outcome-based education and mastery learning. In *Paper presented at the Annual Meeting of the American Educational Research Association*. New Orleans, LA.
Guskey, T. R., & Gates, S. L. (1986). Synthesis of research on the effects of mastery learning in elementary and secondary classrooms. *Educational Leadership*, *May*, 73–80.
Guskey, T. R., & Pigott, T. D. (1988). Research on group-based mastery learning programs: A meta-analysis. *Journal of Educational Research*, *81*(4), 197–216.
Haertel, E. H. (1999). Performance assessment and education reform. *The Phi Delta Kappan*, *80*(9), 662–666.
Hambleton, R. K., Jaeger, R. M., Koretz, D., Linn, R. L., Millman, J., & Phillips, S. E. (1995). Review of the measurement quality of the Kentucky Instructional Results Information System, 1991-1994. Report prepared for the Office of Educational Accountability, Kentucky General Assembly.
Hamilton, L. S. (2003). Assessment as a Policy Tool. *Review of Research in Education*, *27*(1), 25–68. http://doi.org/10.3102/0091732X027001025

Hamilton, L. S., Stecher, B., & Klein, S. P. (2002). *Making Sense of Test-Based Accountability in Education*. Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/pubs/monograph_reports/MR1554.html

Hamilton, L. S., Stecher, B., Marsh, J. A., McCombs, J. S., Robyn, A., Russel, J. L., … Barney, H. (2007). Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states. Santa Monica, CA: RAND Corporation.

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2001). Does peer ability affect student achievement? New York, NY: Andrew W. Mellon Foundation.

Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., & Spurgeon, L. (2015). Student testing in America's great city schools: An inventory and preliminary analysis. Washington, D.C.: Council of the Great City Schools.

Hattie, J. (2009). Visible learning diagram. Retrieved from www.visible-learning.org

Haynes, E., Zeiser, K., Surr, W., Hauser, A., Clymer, L., Walston, J., … Yang, R. (2016). Looking under the hood of competency-based education: The relationship between competency-based education practices and students' learning skills, behaviors, and dispositions. Washington, DC: American Institutes for Research.

Haystead, M. W. (2010). RISC vs. non-RISC schools: A comparison of student proficiencies for reading, writing, and mathematics. Englewood, CO: Marzano Research Laboratory.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlations for planning group randomized experiments in rural education. *Journal of Research in Rural Education*, *22*(10), 1–15.

Herman, J. L. (2004). The effects of testing on instruction. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 141–166). New York, NY: Teachers College Press.

Herman, J. L., & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice*, *12*(4), 20–25.

Herman, J. L., & Linn, R. (2013). On the road to assessing deeper learning: The status of Smarter Balanced and PARCC Assessment Consortia. Los Angeles, CA: CRESST/University of California, Los Angeles.

Herold, D. M., & Fedor, D. B. (2008). *Change the way you lead change: Leadership strategies that really work*. Stanford, CA: Stanford University Press.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177.

Hill, R. K., & DePascale, C. A. (2003). Reliability of No Child Left Behind accountability designs. *Educational Measurement: Issues and Practice*, (Fall), 12–20.

Howe, K. R. (1994). Standards, assessment, and equality of educational opportunity. *Educational Researcher*, *23*(8), 27–33. Retrieved from http://www.jstor.org/stable/1176860

Ingraham, C. (2017). The state with the highest median income is...New Hampshire? Retrieved September 19, 2017, from http://www.pressherald.com/2017/09/13/the-state-with-highest-median-income-is-new-hampshire/

Institute of Education Sciences. (2014). What Works Clearninghouse: Procedures and Standards Handbook (Version 3.0). Washington, DC: Author.

Jellison, J. (2006). *Managing the dynamics of change*. New York, NY: McGraw-Hill.

Jennings, J. L., & Bearak, J. M. (2014). "Teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher*, *43*(8), 381–389. http://doi.org/10.3102/0013189X14554449

Khattri, N., Kane, M. B., & Reeve, A. L. (1995). How performance assessments affect teaching and learning. *Educational Leadership*, *November*, 80–83.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives*, *8*(49), 1–22.
Kliebard, H. M. (2002). Success and failure in educational reform: Are there historical "lessons"? In *Changing course: American curriculum reform in the 20th century* (pp. 126–137). New York, NY: Teachers College Press.
Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores (CSE Report 655). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Koretz, D., Klein, S., McCaffrey, D., & Stecher, B. (1993). Interim report: The reliability of Vermont portfolio scores in the 1992-93 school year (CSE Technical Report 370). Santa Monica, CA: RAND and Center for Research on Evaluation, Standards and Student Testing.
Koretz, D. M., Barron, S., Mitchell, K. J., & Stecher, B. M. (1996). Perceived effects of the Kentucky Instructional Results Information System (KIRIS). Santa Monica, CA: RAND.
Koretz, D. M., Mitchell, K., Barron, S., & Keith, S. (1996). Final report: Perceived effects of the Maryland School Performance Assessment Program (CSE Technical Report 409). *Education*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Koretz, D. M., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice*, *13*(3), 5–16.
Koretz, D. M., Stecher, B., Klein, S., Mccaffrey, D., & Deibert, E. (1993). Can portfolios assess student performance and influence Iistruction? The 1991-1992 Vermont experience. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1992). The reliability of scores from the 1992 Vermont portfolio assessment program: Interim report. CSE Technical Report 355. Santa Monica, CA: RAND Institute on Education and Training.
Kulik, C. L., Kulik, J. A., & Bangert-Drowns, R. L. (1986). Effects of testing for mastery on student learning. In *American Educational Research Association*. San Franscisco, CA.
Kulik, C. L., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research*, *60*, 265–299.
Kulik, J. A., Kulik, C. L., & Bangert-Drowns, R. L. (1990). Is there better evidence on mastery learning? A response to Slavin. *Review of Educational Research*, *60*(2), 303–307.
Kulik, J. A., Kulik, C. L., & Cohen, P. A. (1979). A meta-analysis of outcome studies of Keller's Personalized System of Instruction. *American Psychologist*, *34*, 307–318.
Ladd, H. F., & Goertz, M. E. (Eds.). (2015). *Handbook of Research in Education Finance and Policy* (2nd ed.). New York, NY: Routledge.
Lane, S., Parke, C. S., & Stone, C. A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment*, *8*(4), 279–315. http://doi.org/10.1207/S15326977EA0804
Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed, pp. 387–431). Westport, CT: American Council on Education and Praeger Publishers.
Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.
Le, C., Wolfe, R. E., & Steinberg, A. (2014). The past and the promise: Today's competency education movement. Students at the Center: Competency Education Research Series. Boston, MA: Jobs for the Future.

Lewis, M. W., Eden, R., Garber, C., Rudnick, M., Santibanez, L., & Tsai, T. (2014). Equity in competency education: Realizing the potential, overcoming the obstacles. Boston, MA: Jobs for the Future.

Linn, R. L. (2008). Educational accountability systems. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 3–24). New York, NY: Routledge.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, *31*(6), 3–16.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15–21.

Lyons, S., Evans, C. M., Marion, S. F., & Thompson, J. (2017). New Hampshire Performance Assessment of Competency Education (PACE) Technical Manual (2016-2017). Dover, NH: National Center for the Improvement of Educational Assessment.

Marion, S. F. (2011). Considerations regarding accountability uses of benchmark computer adaptive tests. Dover, NH: National Center for the Improvement of Educational Assessment.

Marion, S., & Leather, P. (2015). Assessment and accountability to support meaningful learning. *Education Policy Analysis Archives*, *23*(9). Retrieved from http://dx.doi.org/10.14507/epaa.v23.1984

McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis*, *17*(3), 305–322.

McMurrer, J. (2007). Choices, changes, and challenges: Curriculum and Instruction in the NCLB era. Washington, DC: Center on Education Policy.

Moss, P. A. (2008). Sociocultural Implications for assessment I: Classroom assessment. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, Equity, and Opportunity to Learn* (pp. 222–258). Cambridge, UK: Cambridge University Press. http://doi.org/10.1017/CBO9780511802157

National Center for Education Statistics. (2015). 2015 mathematics & reading assessments: National scores by student group. Retrieved October 16, 2017, from https://www.nationsreportcard.gov/reading_math_2015/#reading/groups?grade=8

National Center for Education Statistics. (2017). The condition of education 2017: Mathematics performance. Washington, DC: Institute of Education Sciences. Retrieved from https://nces.ed.gov/programs/coe/pdf/coe_cnc.pdf

National Research Council. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. Washington, DC: National Academies Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: The National Academies Press.

National Research Council. (2012). *Education for life and work: developing transferable knowledge and skills in the 21st century*. (J. W. Pellegrino & M. L. Hilton, Eds.). Washington, DC: National Academies Press. http://doi.org/0-309-25649-6

New Hampshire Department of Education. (n.d.). State model competencies.

New Hampshire Department of Education. (2005). High school leadership: Preliminary report. Concord, NH: Author.

New Hampshire Department of Education. (2014a). New Hampshire: Our story of transformation. Concord, NH: Author. Retrieved from http://www.pixar.com/about/Our-Story

New Hampshire Department of Education. (2014b). New Hampshire Performance Assessment of Competency Education: An accountability pilot proposal to the United States Department of Education. Concord, NH: Author.

New Hampshire Department of Education. (2014c). New Hampshire work-study practices.

New Hampshire Department of Education. (2014d). State model competencies. Retrieved June 17, 2015, from http://education.nh.gov/innovations/hs_redesign/competencies.htm
New Hampshire Department of Education. (2015a). NH PACE progress report (April 30, 2015). Concord, NH: Author.
New Hampshire Department of Education. (2015b). Performance Assessment of Competency Education (PACE). Retrieved May 22, 2015, from http://education.nh.gov/assessment-systems/pace.htm
New Hampshire Department of Education. (2015c, March 5). Press Release: Governor Hassan, Department of Education Announce Federal Approval of New Hampshire's Pilot. Retrieved May 22, 2015, from http://education.nh.gov/news/pace.htm
New Hampshire Department of Education. (2016a). Application for inclusion in Performance Assessment for Competency Education PACE 2016-2017. Concord, NH: Author.
New Hampshire Department of Education. (2016b). Moving from good to great in New Hampshire: Performance Assessment of Competency Education (PACE). Concord, NH: Author.
OGAP Math. (n.d.). The Ongoing Assessment (OGAP) Project. Retrieved November 29, 2016, from http://www.ogapmath.com
Pace, L., Moyer, J., & Williams, M. (2015). Building consensus and momentum: A policy and political landscape for K-12 competency education. Cincinnati, OH: KnowledgeWorks.
Pace, L., & Worthen, M. (2014). Laying the foundation for competency education: A policy guide for the next generation educator workforce.
Pane, J. F., Steiner, S. D., Baired, M. D., & Hamilton, L. S. (2015). Continued progress: Promising evidence on personalized learning. Seattle, WA: RAND Corporation and Bill & Melinda Gates Foundation.
Parke, C. S., & Lane, S. (2008). Examining Alignment Between State Performance Assessment and Mathematics Classroom Activities. *The Journal of Educational Research*, *101*(3), 132–147. http://doi.org/10.3200/JOER.101.3.132-147
Parke, C. S., Lane, S., & Stone, C. A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation*, *12*(3), 239–269. http://doi.org/10.1080/13803610600696957
Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers. Chestnut Hill, MA: National Board on Educational Testing and Public Policy.
Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
Pink, D. H. (2009). *Drive: The surprising truth about what motivates us*. New York, NY: Riverhead Books.
Pizmony-Levy, O., & Green Saraisky, N. (2016). Who opts out and why? Results from a national survey on opting out of standardized tests. Research report. New York, NY: Teachers College, Columbia University.
Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan*, *68*, 679–682.
Popham, W. J., Cruse, K. L., Rankin, S. C., Sandifer, P. D., & Williams, P. L. (1985). Measurement-driven instruction: It's on the road. *Phi Delta Kappan*, *66*, 628–634.
Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. O'Connor (Eds.), *Evaluation in Education and Human Services* (pp. 37–75). Netherlands: Springer. http://doi.org/10.1007/978-94-011-2968-8_3

Rogoff, B. (2003). *The cultural nature of human development*. New York, NY: Oxford University Press.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.
Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38.
Rothman, R., & Marion, S. F. (2016). The next generation of state assessment and accountability. *Phi Delta Kappan*, *97*(8), 34–37.
Ryan, S., & Cox, J. D. (2017). Investigating Student Exposure to Competency-Based Education. *Education Policy Analysis Archives*, *25*(24), 1–32. http://doi.org/10.14507/epaa.25.2792
Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262–274. http://doi.org/10.1037/0033-2909.124.2.262
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
Shaw, E. J. (2015). An SAT validity primer. Princeton, NJ: The College Board.
Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14.
Shepard, L. A., & Dougherty, K. C. (1991). Effects of high-stakes testing on instruction. In *Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991)*.
Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V., & Weston, T. J. (1995). Effects of introducting classroom performance assessments on student learning (CSE Technical Report 394). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, *15*(9), 5–11.
Slavin, R. E. (1987a). Mastery learning reconsidered. *Review of Educational Research*, *57*(2), 175–213.
Slavin, R. E. (1987b). Taking the mystery out of mastery: A response to Guskey, Anderson, and Burns. *Review of Educational Research*, *57*(2), 231–235. http://doi.org/10.3102/00346543057002231
Slavin, R. E. (1990). Mastery learning re-reconsidered. *Review of Educational Research*, *60*(2), 300–302.
Smarter Balanced Assessment Consortium. (n.d.). Reporting scores. Retrieved February 21, 2017, from http://www.smarterbalanced.org/assessments/scores/
Smarter Balanced Assessment Consortium. (2015). Smarter balanced assessment consortium: 2013-14 technical report. Author.
Smith, M. L., Noble, A. J., Heinecke, W., Seck, M., Parish, C., Cabay, M., … Bradshaw, A. (1997). Reforming schools by reforming assessment: Consequences of the Arizona student assessment program (ASAP): Equity and teacher capacity building (CSE Technical Report 425). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Smith, M. S., & O'Day, J. (1991). Putting the pieces together: Systemic school reform. *CPRE Policy Briefs*, *6*(4), 1–10.
Spady, W. G. (1977). Competency based education: A bandwagon in search of a definition. *Educational Researcher*, *6*(1), 9–14. http://doi.org/10.3102/0013189X006001009
Spady, W. G. (1978). The concept and implications of competency-based education. *Educational Leadership*, *Oct*, 16–22.
Spady, W. G., & Mitchell, D. E. (1977). Competency based education: Organizational issues and implications. *Educational Researcher*, *6*(2), 9–15.

Stecher, B. (2010). Performance assessment in an era of standards-based accountability. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.
Stecher, B., Barron, S., Kaganoff, T., & Goodwin, J. (1998). The effects of standards-based assessment on classroom practices: Results of the 1996-1997 RAND survey of Kentucky teachers of mathematics and writing (CSE Technical Report 482). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). http://doi.org/10.1017/CBO9781107415324.004
Stecher, B., Barron, S. L., Chun, T., & Ross, K. (2000). The effects of the Washington State education reform on schools and classrooms (CSE Technical Report 525). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Stecher, B., & Chun, T. (2001). The effects of the Washington education reform on school and classroom practice, 1999-2000. Boston, MA: RAND Education.
Stecher, B., Epstein, S., Hamilton, L. S., Marsh, J. A., Robyn, A., McCombs, J. S., … Naftel, S. (2008). *Pain and Gain: Implementing No Child Left Behind in Three States, 2004-2006*. Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org
Stecher, B., & Mitchell, K. J. (1995). Portfolio-driven reform: Vermont teachers' understanding of mathematical problem solving and related changes in classroom practice (CSE Technical Report 400). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Steele, J. L., Lewis, M. W., Santibañez, L., Faxon-Mills, S., Rudnick, M., Stecher, B. M., & Hamilton, L. S. (2014). Competency-based education in three pilot programs: Examining implementation and outcomes. Santa Monica, CA: RAND Education.
Stiggins, R. (2006). Balanced assessment systems: Redefining excellence in assessment. Portland, OR: Educational Testing Service.
Stone, C. A., & Lane, S. (2003). Consequences of a State Accountability Program: Examining Relationships Between School Performance Gains and Teacher, Student, and School Variables. *Applied Measurement in Education*, *16*(1), 1–26. http://doi.org/10.1207/S15324818AME1601_1
Sturgis, C. (2016). Reaching the tipping point: Insights on advancing competency education in New England. Alexandria, VA: International Association for K-12 Online Learning.
Sturgis, C. (2016). Updated: Competency-based education across America. Retrieved from http://www.competencyworks.org/resources/competency-based-education-across-america/
Supovitz, J. (2009). Can high stakes testing leverage educational improvement? Prospects from the last decade of testing and accountability reform. *Journal of Educational Change*, *10*(2–3), 211–227. http://doi.org/10.1007/s10833-009-9105-2
The Education Trust. (2016). New school accountability systems in the states. Retrieved from https://edtrust.org/resource/new-school-accountability-systems-in-the-statesboth-opportunities-and-peril/
The National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. Washington, DC.
Tung, R., & Stazesky, P. (2010). Including performance assessments in accountability systems: A review of scale-up efforts. Boston, MA: Center for Collaborative Education.
U.S. Census Bureau. (2015). State and County QuickFacts: New Hampshire.
U.S. Department of Education. (2015). Fact sheet: Testing action plan. Washington, D.C.: Author. Retrieved from https://www.ed.gov/news/press-releases/fact-sheet-testing-action-plan
Vygotsky, L. S. (1978). *Mind in society*. Cambridge, UK: Cambridge University Press.
Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership*, *May*, 26–33.
Willent, J., Yamashita, J., & Anderson, R. (1983). A meta-analysis of instructional systems applied in science teaching. *Journal of Research in Science Teaching*, *20*(5), 405–417.

Worthen, M., & Pace, L. (2014). A K-12 Federal Policy Framework for Competency Education: Building Capacity for Systems Change, (February), 37. Retrieved from http://www.competencyworks.org/2013/05/how-states-are-advancing-competency-education/

Zeiser, K. L., Taylor, J., Rickles, J., Garet, M. S., & Segeritz, M. (2014). Evidence of deeper learning outcomes. Report #3 Findings from the Study of Deeper Learning: Opportunities and Outcomes. Washington, DC: American Institutes for Research.

APPENDICES

Appendix A Institutional Review Board Approval Not Needed For This Study

According to the University of New Hampshire Institutional Review Board (IRB), use for research purposes of publicly available or anonymous secondary or existing data derived from people does not constitute involving human subjects in research, and thus IRB approval was not needed for this study (http://www.unh.edu/research/sites/www.unh.edu.research/files/docs/RIS/activities_involving_hsirb_approval.pdf).

Appendix B Propensity Score Model

Table B. 1 Parameter Estimates from Propensity Score Model (N=21,632)

	B	S.E.	Wald	df	Sig.	Exp(B)
dpctiep_necap	.081	.010	68.495	1	.000	1.084
dpctfrl_necap	-.001	.003	.215	1	.643	.999
dpctlep_necap	.728	.027	740.204	1	.000	2.071
dpctnonwhite_necap	-.189	.009	475.906	1	.000	.828
dpctmathprof_necap	.009	.006	2.313	1	.128	1.009
dpctELAprof_necap	-.119	.008	207.876	1	.000	.888
Constant	5.366	.592	82.269	1	.000	213.923

Table B.2 Descriptive Statistics on Predicted Probabilities (N=21,632)

Mean	.0746117
Median	.0409281
Std. Deviation	.10270933
Range	.93010
Minimum	.00087
Maximum	.93096

Table B.3 Descriptive Statistics on Predicted Probabilities by Treatment Status

	N	Min	Max	Mean	SD
Non-PACE	20018	.00087	.93096	.0658120	.09560432
PACE	1614	.00317	.82064	.1837517	.12313197

Table B.4 Histogram of Predicted Probabilities by Treatment Status

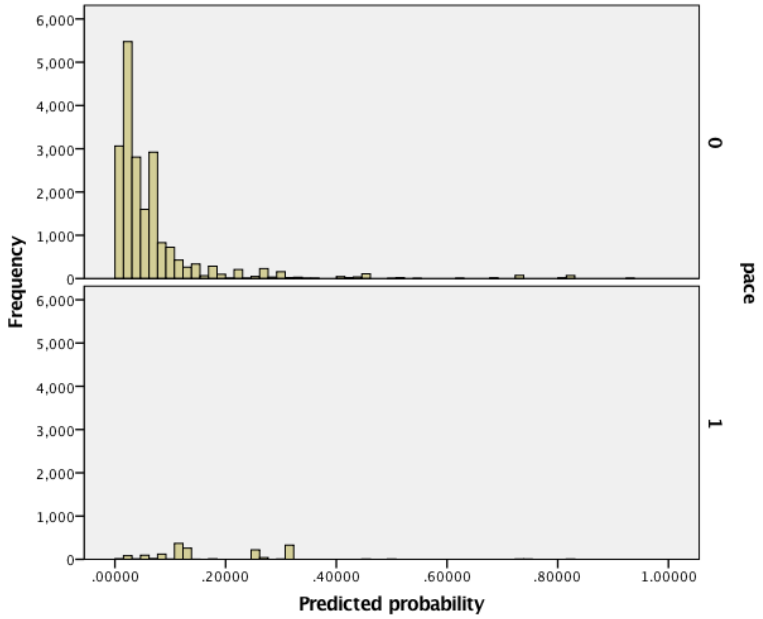


Table B.5 Descriptive Statistics on ATE and ATECV (Corrected Version) Inverse Propensity Score Weighted Variables

		ATE	ATECV
N	Valid	21632	21632
	Missing	0	0
Mean		1.7671	.2043
Median		1.0433	.0687
Std. Deviation		4.79013	.89504
Range		314.38	57.89
Minimum		1.00	.07
Maximum		315.38	57.95

Appendix C Descriptive Statistics for Grade 8 Math by District and Treatment Year in the Unweighted and Weighted Samples

Table C.1 Descriptive statistics for Grade 8 math students in PACE schools by year for unweighted sample

		2014-15 School Year						2015-16 School Year					
District ID							non						non
School ID		sbac	necap	male	iep	frl	white	sbac	necap	male	iep	frl	white
165	N	79	79	79	79	79	79	54	54	54	54	54	54
26505	M	2563.03	648.03	0.39	0	0.32	0.04	2610.98	649.02	0.5	0.09	0.26	0.02
	SD	89.81	10.158	0.491	0	0.468	0.192	82.199	10.117	0.505	0.293	0.442	0.136
461	N	283	283	283	283	283	283	279	279	279	279	279	279
22705	M	2527.67	644.51	0.49	0.18	0.4	0.07	2540.43	642.57	0.48	0.18	0.43	0.06
	SD	101.131	13.743	0.501	0.388	0.491	0.263	105.456	12.314	0.5	0.384	0.496	0.246
476	N	115	115	115	115	115	115	123	123	123	123	123	123
20630	M	2535.7	646.36	0.39	0.15	0.19	0.03	2591.8	646.17	0.55	0.11	0.18	0.06
	SD	93.599	11.756	0.49	0.356	0.395	0.16	102.769	9.41	0.499	0.309	0.385	0.233
111	N	321	321	321	321	321	321	241	241	241	241	241	241
20270	M	2582.81	646.73	0.48	0.15	0.34	0.16	2603.82	645.15	0.54	0.11	0.32	0.16
	SD	109.093	13.207	0.5	0.36	0.473	0.366	103.5	12.262	0.5	0.316	0.466	0.365
365	N	8	8	8	8	8	8	5	5	5	5	5	5
20885	M	2609.75	647.63	0.38	0	0.13	0	2542.8	646.6	0	0	0.4	0
	SD	83.998	8.634	0.518	0	0.354	0	85.491	7.127	0	0	0.548	0
439	N	40	40	40	40	40	40	44	44	44	44	44	44
26550	M	2514.2	643.03	0.63	0.15	0.63	0.03	2496.73	639.27	0.45	0.23	0.45	0.02
	SD	102.593	11.146	0.49	0.362	0.49	0.158	96.854	11.884	0.504	0.424	0.504	0.151
705	N	12	12	12	12	12	12	10	10	10	10	10	10
28400	M	2558.42	647	0.5	0.08	0.08	0	2622	648.9	0.3	0	0.2	0
	SD	55.884	7.954	0.522	0.289	0.289	0	63.871	9.073	0.483	0	0.422	0

Note. Red highlights indicate no treatment; green highlights indicate one year of treatment; and orange highlights indicate two years of treatment.

Table C.2 Descriptive statistics for Grade 8 math students in PACE schools by treatment year in the weighted sample

District ID / *School ID*		2014-15 School Year						2015-16 School Year					
		sbac	necap	male	iep	frl	non white	sbac	necap	male	iep	frl	non white
165	N	1519	1519	1519	1519	1519	1519	1749	1749	1749	1749	1749	1749
26505	M	2562.35	647.52	0.39	0	0.34	0.03	2613.12	649.31	0.51	0.09	0.26	0.02
	SD	85.565	10.471	0.489	0	0.473	0.182	81.867	10.147	0.5	0.282	0.44	0.138
461	N	2951	2951	2951	2951	2951	2951	2735	2735	2735	2735	2735	2735
22705	M	2520.75	644.15	0.57	0.22	0.43	0.06	2531.13	641.82	0.56	0.2	0.51	0.06
	SD	96.252	12.497	0.495	0.416	0.496	0.235	108.82	11.879	0.496	0.403	0.5	0.241
476	N	1472	1472	1472	1472	1472	1472	1528	1528	1528	1528	1528	1528
20630	M	2532.7	646.24	0.4	0.16	0.21	0.03	2596.32	645.78	0.44	0.08	0.34	0.04
	SD	94.735	11.614	0.49	0.37	0.406	0.158	95.981	8.476	0.496	0.266	0.473	0.199
111	N	1340	1340	1340	1340	1340	1340	1491	1491	1491	1491	1491	1491
20270	M	2572.8	646.87	0.5	0.14	0.33	0.13	2604.56	644.63	0.53	0.14	0.41	0.1
	SD	105.152	11.808	0.5	0.348	0.469	0.331	95.603	12.401	0.499	0.347	0.492	0.305
365	N	33	33	33	33	33	33	7	7	7	7	7	7
20885	M	2658.34	654.07	0.18	0	0.06	0	2542.8	646.6	0	0	0.4	0
	SD	72.704	8.463	0.39	0	0.241	0	82.946	6.915	0	0	0.531	0
439	N	206	206	206	206	206	206	346	346	346	346	346	346
26550	M	2495.95	644.48	0.73	0.11	0.73	0.02	2495.89	639.56	0.53	0.38	0.49	0.01
	SD	108.091	9.705	0.443	0.309	0.447	0.132	90.647	9.081	0.5	0.485	0.501	0.095
705	N	637	637	637	637	637	637	133	133	133	133	133	133
28400	M	2563.09	647.24	0.29	0.06	0.49	0	2631.81	651.93	0.45	0	0.11	0
	SD	40.312	6.248	0.455	0.243	0.5	0	60.037	6.109	0.499	0	0.311	0

Note. Red indicates no treatment; green indicates one year of treatment; and orange indicates two years of treatment.

Appendix D Taxonomies of Multi-Level Models used to Select the "Final" Grade 8 Math Models Shown in Table 4.3

Table D.1 Parameter estimates and goodness of fit statistics from a taxonomy of M1 models showing the effects of student-level characteristics on Grade 8 math achievement using inverse propensity score weights

	Model 0		**Model 1a**		**Model 1b**		**Model 1c**		**Model 1d**	
	B	SE	B	SE	B	SE	B	SE	B	SE
Intercept	2577.47***	3.56	2575.48***	2.13	2578.87***	2.10	2580.74***	2.13	2587.11***	2.15
necap			6.54***	0.04	6.44***	0.04	6.26***	0.04	6.26***	0.04
frl					-12.48***	0.98	-12.05***	0.98	-12.49***	0.98
iep							-14.29***	1.37	-11.38***	1.38
male									-13.18***	0.85
Random Effects										
σ^2	16605.63***	160.13	6816.35***	65.73	6767.38***	65.26	6732.66***	64.93	6659.02***	64.22
τ_{00}	1335.36***	199.78	470.34***	71.91	449.11***	69.25	460.63***	70.80	449.39***	69.14
%Reduction σ^2			0.59		0.59		0.59		0.60	
%Reduction τ_{00}			0.65		0.66		0.66		0.66	
Goodness of fit										
-2 LL	267238.75		247961.19		247801.20		247693.07		247453.74	
AIC	267244.75		247969.19		247811.20		247705.07		247467.74	
BIC	267268.70		248001.12		247851.11		247752.96		247523.62	

~p<.10, *p<.05, **p<.01, ***p<.001

Note. MIXED command in SPSS with ML estimation; inverse propensity score weights applied as a regression weight. B=unstandardized parameter coefficient; SE=standard error; AIC=Akaike's Information Criteria; BIC=Schwarz's Bayesian Criterion.

Table D.2 Parameter estimates and goodness of fit statistics from a taxonomy of M2 means as outcomes models showing the effects of school-level characteristics on Grade 8 math achievement using inverse propensity score weights

	Model 2a		Model 2b		Model 2c		Model 2d		Model 2e		Model 2f	
Variables	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE
Intercept	2577.56***	3.08	2577.44***	3.15	2576.51***	2.65	2573.42***	2.68	2570.21***	2.71	2570.31***	2.71
pctmathprof	0.95***	0.09	0.96***	0.09	0.91***	0.09	0.96***	0.09	1.12***	0.09	1.11***	0.09
pctiep			1.38***	0.30	1.24***	0.29	1.46***	0.29	0.47	0.29		
pctfrl					-1.14***	0.16	-1.19***	0.16	-1.01***	0.16	-1.01***	0.15
Nkids							-0.05***	0.01	-0.02*	0.01	-0.02*	0.01
sbacid									11.13***	1.75	11.14***	1.75
treat1									-32.04*	14.54	-32.86*	14.51
treat2									-7.85	14.53	-7.80	14.52
sbacid*treat1									42.82*	15.44	43.44*	15.43
Variance components												
σ^2	16556.26***	159.68	16536.23***	159.49	16532.06***	159.46	16483.24***	158.98	16292.93***	157.15	16295.03***	157.18
τ_{00}	965.44***	154.18	1014.28***	160.59	677.49***	114.56	676.85***	113.40	594.84***	103.17	593.56***	103.38
%Reduction τ_{00}	0.28		0.24		0.49		0.49		0.55		0.56	
Goodness of fit												
-2LL	267141.013		267119.982		267074.354		267010.565		266748.096		266750.68	
AIC	267149.013		267129.982		267086.354		267024.565		266770.096		266770.68	
BIC	267180.941		267169.892		267134.246		267080.438		266857.898		266850.50	

~p<.10, *p<.05, **p<.01, ***p<.001

Note. MIXED command in SPSS with ML estimation; inverse propensity score weights applied as a regression weight. B=unstandardized parameter coefficient; SE=standard error; AIC=Akaike's Information Criteria; BIC=Schwarz's Bayesian Criterion.

Table D.3 Parameter estimates and goodness of fit statistics from a taxonomy of M3 combined level-1 and level-2 models showing the effects of student- and school-level characteristics on Grade 8 math achievement using inverse propensity score weights

	Model 3= M1d + M2f	
Variables	B	SE
Intercept	2578.55***	2.19
necap	6.35***	0.04
frl	-12.79***	0.95
iep	-9.93***	1.34
male	-13.72***	0.83
pctmathprof	-0.38***	0.06
pctfrl	-0.52***	0.12
Nkids	-0.02***	0.01
sbacid	13.83***	1.09
treat1	-30.44*	12.08
treat2	-4.31	12.08
sbacid*treat1	43.90***	12.56
Variance Components		
σ^2	6277.09***	60.54
τ_{00}	417.14***	64.53
%Reduction σ^2	0.62	
%Reduction τ_{00}	0.69	
Goodness of fit		
-2LL	246174.456	
AIC	246202.456	
BIC	246314.203	

~p<.10, *p<.05, **p<.01, ***p<.001

Note. MIXED command in SPSS with ML estimation; inverse propensity score weights applied as a regression weight. B=unstandardized parameter coefficient; SE=standard error; AIC=Akaike's Information Criteria; BIC=Schwarz's Bayesian Criterion.

Table D.4 Parameter estimates and goodness of fit statistics from a taxonomy of M4 cross-level effect models showing the effects of student- and school-level characteristics on Grade 8 math achievement using inverse propensity score weights

	Model 4a		Model 4b		Model 4c		Model 4d	
Variables	B	SE	B	SE	B	SE	B	SE
Intercept	2578.59***	2.20	2579.12***	2.21	2579.99***	2.23	2576.92***	2.25
necap	6.41***	0.05	6.39***	0.05	6.31***	0.05	6.29***	0.05
frl	-12.63***	0.95	-14.49***	1.28	-14.42***	1.27	-14.24***	1.27
iep	-9.46***	1.34	-9.55***	1.34	-15.44***	1.71	-16.75***	1.71
male	-14.10***	0.82	-14.08***	0.82	-14.12***	0.82	-7.86***	1.02
pctmathprof	-0.35***	0.06	-0.37***	0.06	-0.36***	0.06	-0.34***	0.06
pctfrl	-0.42***	0.12	-0.41***	0.12	-0.41***	0.12	-0.42***	0.12
Nkids	-0.02***	0.01	-0.02***	0.01	-0.02***	0.01	-0.02***	0.01
sbacid	14.05***	1.09	13.97***	1.09	13.91***	1.09	13.91***	1.08
treat1	-30.90*	12.10	-32.26*	12.15	-33.34*	12.22	-25.42*	12.30
treat2	-4.06	12.11	-5.43	12.16	-8.22	12.23	0.33	12.33
sbacid*treat1	44.22***	12.59	43.81***	12.61	44.03***	12.67	44.53***	12.71
treat1*necap	-0.65***	0.09	-0.62***	0.09	-0.51***	0.10	-0.54***	0.10
treat2*necap	0.68***	0.10	0.72***	0.11	1.06***	0.12	1.08***	0.12
treat1*frl			4.26	2.28	3.93	2.30	3.02	2.29
treat2*frl			4.16	2.54	3.80	2.54	3.16	2.53
treat1*iep					9.01*	3.21	13.63***	3.26
treat2*iep					23.99***	3.76	27.79***	3.78
treat1*male							-17.49***	2.11
treat2*male							-17.68***	2.30
Variance components								
σ^2	6239.97***	60.18	6238.46***	60.16	6225.84***	60.04	6195.46***	59.75
τ_{00}	419.10***	64.98	420.83***	65.24	425.31***	65.92	428.49***	66.26
%Reduction σ^2	0.62		0.62		0.63		0.63	
%Reduction τ_{00}	0.69		0.68		0.68		0.68	
Goodness of fit								
-2LL	246047.245		246042.464		245999.967		245895.422	
AIC	246079.245		246078.464		246039.967		245939.422	
BIC	246206.956		246222.138		246199.606		246115.024	

Appendix E Sensitivity Analysis of Treatment Effects to Weighting in Grade 8 Math

Table E.1 Parameter estimates and goodness of fit statistics from selected multi-level models showing the effects of student- and school-level characteristics on Grade 8 math achievement for the unweighted sample

	M0: Null		M1: Level-1 Only		M2: Level-2 Only		M3: Levels 1&2		M4: Cross-Level	
Variables	B	SE	B	SE	B	SE	B	SE	B	SE
Intercept	2577.72***	3.34	2589.24***	2.08	2570.62***	2.66	2578.48***	2.42	2578.06***	2.44
necap			6.30***	0.04			6.36***	0.04	6.36***	0.04
frl			-14.31***	1.03			-13.89***	1.03	-13.98***	1.03
iep			-15.59***	1.36			-14.61***	1.35	-15.73***	1.39
male			-8.52***	0.82			-8.71***	0.81	-7.96***	0.83
pctmathprof					1.95***	0.10	-0.09	0.06	-0.09	0.06
pctfrl					-0.71***	0.14	-0.21	0.12	-0.20	0.12
Nkids					-0.02*	0.01	-0.02***	0.01	-0.02***	0.01
sbacid					12.58***	1.40	14.54***	0.86	14.54***	0.86
treat1					-29.33*	13.57	-30.59*	12.30	-27.80*	12.52
treat2					-8.78	13.65	-9.23	12.34	-2.88	12.74
sbacid*treat1					21.92	15.68	26.68*	13.38	26.86*	13.45
treat1*necap									-0.43*	0.19
treat2*necap									0.83***	0.28
treat1*iep									19.93*	7.17
treat2*iep									22.21*	9.07
treat1*male									-13.38***	4.38
treat2*male									-15.80*	5.62
Variance components										
σ^2	9451.91***	91.15	3528.60***	34.03	9274.65***	89.46	3455.37***	33.33	3448.00***	33.26
τ_{00}	1203.77***	178.49	440.07***	66.21	460.47***	78.70	411.43***	63.25	415.92***	64.03
%Reduction σ^2			0.63		0.02		0.63		0.64	
%Reduction τ_{00}			0.63		0.62		0.66		0.65	
Goodness of fit										
-2LL	259738.95		238422.35		259232.96		237963.74		237918.94	
AIC	259744.95		238436.35		259252.96		237991.74		237958.94	
BIC	259768.89		238492.22		259332.78		238103.48		238118.58	

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

Note. MIXED command in SPSS with ML estimation; B=unstandardized parameter coefficient; SE=standard error; AIC = Akaike's Information Criteria; BIC = Schwarz's Bayesian Criterion.

Appendix F Descriptive Statistics for Grade 8 ELA by District and Treatment Year in the Unweighted and Weighted Samples

Table F.1 Descriptive statistics for Grade 8 ELA students in PACE schools by year for unweighted sample

		2014-15 School Year						2015-16 School Year					
District ID							non						non
School ID		sbac	necap	male	iep	frl	white	sbac	necap	male	iep	frl	white
165	N	78	79	79	79	79	79	54	54	54	54	54	54
26505	M	2579.50	651.19	0.39	0.00	0.32	0.04	2604.46	655.15	0.50	0.09	0.26	0.02
	SD	81.48	13.84	0.49	0.00	0.47	0.19	74.09	13.23	0.51	0.29	0.44	0.14
461	N	269	282	282	282	282	282	275	279	279	279	279	279
22705	M	2563.61	647.76	0.49	0.18	0.40	0.07	2558.18	643.96	0.48	0.18	0.43	0.06
	SD	88.99	12.55	0.50	0.39	0.49	0.26	88.53	11.08	0.50	0.38	0.50	0.25
476	N	114	115	115	115	115	115	123	123	123	123	123	123
20630	M	2528.26	647.84	0.39	0.15	0.19	0.03	2611.33	648.46	0.55	0.11	0.18	0.06
	SD	92.18	11.72	0.49	0.36	0.40	0.16	73.16	10.03	0.50	0.31	0.39	0.23
111	N	320	321	321	321	321	321	240	240	240	240	240	240
20270	M	2577.41	649.76	0.48	0.15	0.34	0.16	2598.62	646.88	0.53	0.11	0.31	0.15
	SD	87.23	12.91	0.50	0.36	0.47	0.37	89.08	12.77	0.50	0.32	0.46	0.36
365	N	8	8	8	8	8	8	5	5	5	5	5	5
20885	M	2609.00	651.13	0.38	0.00	0.13	0.00	2615.60	651.60	0.00	0.00	0.40	0.00
	SD	57.12	16.47	0.52	0.00	0.35	0.00	49.70	11.48	0.00	0.00	0.55	0.00
439	N	39	40	40	40	40	40	44	44	44	44	44	44
26550	M	2548.00	645.07	0.63	0.15	0.63	0.03	2511.70	642.05	0.45	0.23	0.45	0.02
	SD	73.71	10.59	0.49	0.36	0.49	0.16	80.13	11.34	0.50	0.42	0.50	0.15
705	N	12	12	12	12	12	12	10	10	10	10	10	10
28400	M	2598.92	647.17	0.50	0.08	0.08	0.00	2649.10	658.30	0.30	0.00	0.20	0.00
	SD	80.01	9.85	0.52	0.29	0.29	0.00	30.15	10.09	0.48	0.00	0.42	0.00

Note. Red highlights indicate no treatment; green highlights indicate one year of treatment; and orange highlights indicate two years of treatment.

Table F.2 Descriptive statistics for Grade 8 ELA students in PACE schools by treatment year in the weighted sample

District ID / *School ID*		2014-15 School Year						2015-16 School Year					
		sbac	necap	male	iep	frl	non white	sbac	necap	male	iep	frl	non white
165	N	1496	1513	1513	1513	1513	1513	1744	1744	1744	1744	1744	1744
26505	M	2577.42	651.35	0.39	0.00	0.34	0.03	2606.03	655.14	0.51	0.09	0.26	0.02
	SD	77.23	13.82	0.49	0.00	0.47	0.18	73.93	13.27	0.50	0.28	0.44	0.14
461	N	2841	2957	2957	2957	2957	2957	2721	2750	2750	2750	2750	2750
22705	M	2557.98	647.47	0.57	0.22	0.43	0.06	2554.09	644.12	0.56	0.20	0.51	0.06
	SD	85.28	11.70	0.50	0.42	0.50	0.23	90.32	11.50	0.50	0.40	0.50	0.24
476	N	1461	1474	1474	1474	1474	1474	1530	1530	1530	1530	1530	1530
20630	M	2526.21	647.82	0.40	0.16	0.21	0.03	2632.60	649.99	0.44	0.08	0.34	0.04
	SD	91.97	11.56	0.49	0.37	0.41	0.16	80.96	9.46	0.50	0.27	0.47	0.20
111	N	1331	1334	1334	1334	1334	1334	1485	1485	1485	1485	1485	1485
20270	M	2576.71	649.86	0.51	0.14	0.33	0.13	2600.98	645.58	0.53	0.14	0.41	0.10
	SD	81.33	12.01	0.50	0.35	0.47	0.33	82.65	11.49	0.50	0.35	0.49	0.30
365	N	33	33	33	33	33	33	7	7	7	7	7	7
20885	M	2633.97	660.94	0.18	0.00	0.06	0.00	2615.60	651.60	0.00	0.00	0.40	0.00
	SD	44.78	14.46	0.39	0.00	0.24	0.00	48.21	11.14	0.00	0.00	0.53	0.00
439	N	202	206	206	206	206	206	346	346	346	346	346	346
26550	M	2554.27	647.29	0.73	0.11	0.73	0.02	2514.17	640.24	0.53	0.37	0.49	0.01
	SD	70.76	9.99	0.44	0.31	0.45	0.13	90.71	11.04	0.50	0.49	0.50	0.09
705	N	635	635	635	635	635	635	132	132	132	132	132	132
28400	M	2630.16	654.37	0.29	0.06	0.50	0.00	2655.61	659.34	0.45	0.00	0.11	0.00
	SD	68.24	10.78	0.45	0.24	0.50	0.00	28.41	9.09	0.50	0.00	0.31	0.00

Note. Red indicates no treatment; green indicates one year of treatment; and orange indicates two years of treatment.

Appendix G Taxonomies of Multi-Level Models used to Select the "Final" ELA Models Shown in Table 4.8

Table G.1 Parameter estimates and goodness of fit statistics from a taxonomy of M1 models showing the effects of student-level characteristics on Grade 8 ELA achievement using inverse propensity score weights

	Model 0		Model 1a		Model 1b		Model 1c		Model 1d		Model 1e	
	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE
Intercept	2592.54***	3.02	2590.17***	2.29	2593.16***	2.24	2596.83***	2.27	2605.36***	2.30	2605.51***	2.30
necap			4.86***	0.04	4.79***	0.04	4.42***	0.04	4.26***	0.04	4.26***	0.04
frl					-11.00***	0.98	-9.89***	0.97	-11.01***	0.97	-10.92***	0.97
iep							-28.28***	1.39	-27.33***	1.38	-27.36***	1.38
male									-16.42***	0.86	-16.44***	0.86
nonwhite											-2.25	1.63
Random Effects												
σ^2	12527.89***	121.45	6760.50***	65.54	6722.74***	65.18	6593.26***	63.92	6483.21***	62.85	6482.64***	62.85
τ_{00}	956.79***	142.25	552.02***	81.73	521.46***	77.90	533.43***	79.13	524.63***	77.89	524.47***	77.87
%Reduction σ^2			0.46		0.46		0.47		0.48		0.48	
%Reduction τ_{00}			0.42		0.45		0.44		0.45		0.45	
Goodness of fit												
-2 LL	258371.12		245176.73		245051.44		244639.69		244279.48		244277.58	
AIC	258377.12		245184.73		245061.44		244651.69		244293.48		244293.58	
BIC	258371.12		245216.62		245101.30		244699.52		244349.28		244357.35	

~p<.10, *p<.05, **p<.01, ***p<.001

Note. MIXED command in SPSS with ML estimation; inverse propensity score weights applied as a regression weight. B=unstandardized parameter coefficient; SE=standard error; AIC=Akaike's Information Criteria; BIC=Schwarz's Bayesian Criterion.

Table G.2 Parameter estimates and goodness of fit statistics from a taxonomy of M2 means as outcomes models showing the effects of school-level characteristics on Grade 8 ELA achievement using inverse propensity score weights

	Model 2a		Model 2b		Model 2c		Model 2d		Model 2e		Model 2f	
	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE
Intercept	2591.18***	2.46	2591.27***	2.46	2590.33***	2.05	2588.44***	2.00	2584.22***	2.23	2583.85***	2.09
pctELAprof	1.38***	0.09	1.40***	0.09	1.31***	0.09	1.32***	0.09	1.41***	0.09	1.38***	0.09
pctiep			-0.88***	0.26	-0.81***	0.25	-0.68*	0.24	-1.72***	0.25		
pctfrl					-0.96***	0.13	-1.02***	0.12	-0.94***	0.13	-0.97***	0.12
Nkids							-0.03***	0.00	-0.03***	0.01	-0.03***	0.01
sbacid									10.73***	1.51	10.55***	1.51
treat1									-33.25*	11.73	-30.01*	10.79
treat2									-7.52	11.72	-7.28	10.79
sbacid*treat1									21.35	12.54	18.96	11.65
Variance components												
σ^2	12423.69***	120.46	12416.85***	120.40	12411.65***	120.38	12394.32***	120.19	12185.46***	118.19	12220.53***	118.50
τ_{00}	597.47***	96.67	596.57***	96.95	386.25***	70.21	351.28***	63.12	385.12***	70.28	324.46***	59.10
%Reduction τ_{00}	0.38		0.38		0.60		0.63		0.60		0.66	
Goodness of fit												
-2LL	258145.21		258133.34		258082.92		258044.50		257690.72		257736.45	
AIC	258153.21		258143.34		258094.92		258058.50		257712.72		257756.45	
BIC	258185.09		258183.19		258142.74		258114.30		257800.40		257836.17	

~p<.10, *p<.05, **p<.01, ***p<.001

Note. MIXED command in SPSS with ML estimation; inverse propensity score weights applied as a regression weight. B=unstandardized parameter coefficient; SE=standard error; AIC=Akaike's Information Criteria; BIC=Schwarz's Bayesian Criterion.

Table G.3 Parameter estimates and goodness of fit statistics from a taxonomy of M3 combined level-1 and level-2 models showing the effects of student- and school-level characteristics on Grade 8 ELA achievement using inverse propensity score weights

	Model 3= M1d + M2f	
	B	SE
Intercept	2597.76***	2.15
necap	4.24***	0.04
frl	-11.45***	0.95
iep	-27.39***	1.35
male	-17.11***	0.85
pctELAprof	0.39***	0.07
pctfrl	-0.58***	0.11
Nkids	-0.04***	0.01
sbacid	10.41***	1.09
treat1	-33.35*	11.71
treat2	-9.44	11.71
sbacid*treat1	33.21*	12.20
Variance Components		
σ^2	6224.13***	60.35
τ_{00}	391.89***	61.56
%Reduction σ^2	0.50	
%Reduction τ_{00}	0.59	
Goodness of fit		
-2LL	243380.60	
AIC	243408.60	
BIC	243520.20	

~p<.10, *p<.05, **p<.01, ***p<.001

Note. MIXED command in SPSS with ML estimation; inverse propensity score weights applied as a regression weight. B=unstandardized parameter coefficient; SE=standard error; AIC=Akaike's Information Criteria; BIC=Schwarz's Bayesian Criterion.

Table G.4 Parameter estimates and goodness of fit statistics from a taxonomy of M4 cross-level effect models showing the effects of student- and school-level characteristics on Grade 8 ELA achievement using inverse propensity score weights

	Model 4a		Model 4b		Model 4c		Model 4d	
Variables	B	SE	B	SE	B	SE	B	SE
Intercept	2597.79***	2.15	2598.73***	2.17	2599.18***	2.17	2596.94***	2.20
necap	4.20***	0.05	4.18***	0.05	4.14***	0.05	4.18***	0.05
frl	-11.47***	0.95	-14.77***	1.27	-14.69***	1.27	-14.39***	1.27
iep	-27.43***	1.35	-27.63***	1.35	-30.80***	1.72	-30.84***	1.71
male	-17.11***	0.85	-17.08***	0.85	-17.09***	0.85	-12.89***	1.05
pctELAprof	0.40***	0.07	0.38***	0.07	0.38***	0.07	0.38***	0.07
pctfrl	-0.58***	0.12	-0.55***	0.12	-0.55***	0.12	-0.56***	0.12
Nkids	-0.04***	0.01	-0.04***	0.01	-0.04***	0.01	-0.04***	0.01
sbacid	10.41***	1.09	10.34***	1.09	10.33***	1.09	10.33***	1.08
treat1	-33.44*	11.71	-37.09***	11.76	-37.39***	11.75	-35.80***	11.84
treat2	-9.51	11.71	-10.48	11.77	-12.51	11.77	-0.57	11.87
sbacid*treat1	33.33*	12.20	32.67*	12.22	32.79*	12.21	32.71*	12.25
treat1*necap	0.07	0.09	0.15	0.09	0.19	0.10	0.15	0.10
treat2*necap	0.14	0.10	0.16	0.10	0.40***	0.11	0.17	0.11
treat1*frl			10.99***	2.31	11.03***	2.32	10.75***	2.32
treat2*frl			3.31	2.50	2.63	2.51	1.58	2.50
treat1*iep					2.22	3.31	2.15	3.32
treat2*iep					17.52***	3.88	18.28***	3.87
treat1*male							-2.87	2.17
treat2*male							-22.80***	2.36
Variance components								
σ^2	6223.46***	60.35	6216.70***	60.28	6210.75***	60.22	6183.17***	59.96
τ_{00}	391.52***	61.55	393.19***	61.89	392.55***	61.78	395.17***	62.01
%Reduction σ^2	0.50		0.50		0.50		0.51	
%Reduction τ_{00}	0.59		0.59		0.59		0.59	
Goodness of fit								
-2LL	243378.20		243355.48		243334.95		243240.83	
AIC	243410.20		243391.48		243374.95		243284.83	
BIC	243537.74		243534.96		243534.37		243460.19	

Appendix H Sensitivity Analysis of Treatment Effects to Weighting in Grade 8 ELA

Table H.1 Parameter estimates and goodness of fit statistics from selected multi-level models showing the effects of student- and school-level characteristics on Grade 8 ELA achievement for the unweighted sample

	M0: Null		M1: Level-1 Only		M2: Level-2 Only		M3: Levels 1&2		M4: Cross-Level	
Variables	B	SE	B	SE	B	SE	B	SE	B	SE
Intercept	2592.72***	2.90	2607.80***	2.20	2584.71***	2.23	2597.56***	2.39	2597.46***	2.40
necap			4.17***	0.04			4.17***	0.04	4.17***	0.04
frl			-17.64***	1.02			-16.59***	1.03	-16.34***	1.07
iep			-32.07***	1.35			-31.91***	1.34	-31.98***	1.37
male			-12.73***	0.83			-12.85***	0.82	-12.74***	0.84
pctELAprof					1.77***	0.10	0.31***	0.07	0.32***	0.07
pctfrl					-0.79***	0.12	-0.43***	0.12	-0.43***	0.12
Nkids					-0.03***	0.01	-0.03***	0.01	-0.03***	0.01
sbacid					11.10***	1.19	10.51***	0.85	10.52***	0.85
treat1					-23.84*	11.38	-31.11*	12.12	-31.97*	12.37
treat2					-6.89	11.44	-10.46	12.15	-4.26	12.63
sbacid*treat1					15.12	13.19	22.61	13.20	22.67	13.20
treat1*frl									0.16	4.71
treat2*frl									-8.84	6.04
treat1*iep									0.38	6.62
treat2*iep									2.58	8.07
treat1*male									1.49	4.39
treat2*male									-7.38	5.57
Variance components										
σ^2	6816.66***	66.09	3433.85***	33.29	6705.25***	65.02	3388.75***	32.86	3388.10***	32.86
τ_{00}	908.14***	131.65	495.47***	71.68	321.10***	54.58	398.39***	60.92	398.68***	60.97
%Reduction σ^2			0.50		0.02		0.50		0.50	
%Reduction τ_{00}			0.45		0.65		0.56		0.56	
Goodness of fit										
-2LL	249977.48		235311.73		249520.26		235006.71		235002.70	
AIC	249983.48		235325.73		249540.26		235034.71		235042.70	
BIC	250007.40		235381.53		249619.97		235146.31		235202.12	

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

Note. MIXED command in SPSS with ML estimation; B=unstandardized parameter coefficient; SE=standard error; AIC = Akaike's Information Criteria; BIC = Schwarz's Bayesian Criterion.