education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

Arizona State University

Volume 27 Number 10 February 4, 2019 ISSN 1068-2341

Effects of New Hampshire's Innovative Assessment and Accountability System on Student Achievement Outcomes After Three Years

Carla M. Evans
National Center for the Improvement of Educational Assessment
United States

Citation: Evans, C. M. (2019). Effects of New Hampshire's innovative assessment and accountability system on student achievement outcomes after three years. *Education Policy Analysis Archives, 27*(10). http://dx.doi.org/10.14507/epaa.27.4014

Abstract: New Hampshire's Performance Assessment of Competency Education (PACE) pilot received a waiver from federal statutory requirements related to state annual achievement testing starting in the 2014-15 school year. PACE is considered an "innovative" assessment and accountability system because performance assessments are used to help determine student proficiency in most federally required grades and subjects instead of the state achievement test. One key criterion for success in the early years of the PACE innovative assessment system is "no harm" on the statewide accountability test. This descriptive study examines the effect of PACE on Grades 8 and 11 mathematics and English language arts student achievement during the first three years of implementation (2014-15, 2015-16, and 2016-17 school years) and the extent to which those effects vary for certain student subgroups using results from the state's accountability tests (Smarter Balanced and SATs). Findings suggest that students in PACE schools tend to exhibit small positive effects on the Grades 8 and 11 state achievement tests in both subjects in comparison to students attending non-PACE comparison schools. Lower achieving students tended to exhibit small positive differential effects, whereas male students tended to exhibit small negative differential effects. Implications for research, policy, and practice are discussed.

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 7/5/2018
Revisions received: 9/10/2018
Accepted: 9/10/2018

Keywords: Performance Based Assessment; Academic Achievement; Accountability; Educational Policy; Hierarchical Linear Modeling

Efectos de la evaluación innovadora y el sistema de rendición de cuentas de New Hampshire en los resultados del rendimiento estudiantil después de tres años

Resumen: La Evaluación de Desempeño de la Educación en Competencias (PACE) de New Hampshire se considera un sistema "innovador" de evaluación y rendición de cuentas porque las evaluaciones de desempeño se usan para ayudar a determinar la competencia de los estudiantes en la mayoría de las calificaciones y materias requeridas por el gobierno federal en lugar de la prueba de rendimiento estatal. Un criterio clave para el éxito de PACE es "no causar daño" en la prueba de responsabilidad estatal. Este estudio descriptivo examina el efecto de PACE en los logros estudiantiles de matemáticas y de artes del lenguaje en inglés en los grados 8 y 11 durante los primeros tres años de implementación (2014-2017) y la medida en que esos efectos varían para ciertos subgrupos de estudiantes utilizando los resultados de la responsabilidad del estado pruebas (*Smarter Balanced* y *SAT*). Los hallazgos sugieren que los estudiantes en las escuelas PACE tienden a exhibir pequeños efectos positivos en los exámenes de logros estatales de los grados 8 y 11 en ambas materias en comparación con los estudiantes que asisten a escuelas comparativas que no pertenecen a PACE. Los estudiantes con bajo rendimiento tendían a exhibir pequeños efectos diferenciales positivos, mientras que los estudiantes varones tendían a exhibir pequeños efectos diferenciales negativos. Se discuten las implicaciones para la investigación, la política y la práctica.

Palabras-clave: Evaluación basada en el desempeño; logro académico; rendición de cuentas; política educativa

Efeitos do sistema inovador de avaliação e *accountability* de New Hampshire nos resultados de desempenho dos alunos após três anos

Resumo: A Avaliação de Competências de Educação de Desempenho de New Hampshire (PACE) é considerada um sistema "inovador" de avaliação e a*ccountability* porque as avaliações de desempenho são usadas para ajudar a determinar a proficiência dos alunos na maioria das as qualificações e assuntos exigidos pelo governo federal em vez do teste de desempenho do estado. Um critério-chave para o sucesso do PACE é "não causar danos" no teste de responsabilidade do estado. Este estudo descritivo examina o efeito do PACE no desempenho do aluno em matemática e artes da língua inglesa nos graus 8 e 11 durante os primeiros três anos de implementação (2014-2017) e a extensão em que esses efeitos variam para determinados subgrupos de alunos usando os resultados dos testes de responsabilidade do estado (Smarter Balanced e SAT). As descobertas sugerem que os alunos das escolas PACE tendem a apresentar pequenos efeitos positivos nos testes de desempenho estadual nas 8ª e 11ª séries dos dois grupos, em comparação com os alunos que frequentam escolas comparativas que não pertencem ao PACE. Estudantes com baixo desempenho tendem a exibir pequenos efeitos diferenciais positivos, enquanto os estudantes do sexo masculino tendem a exibir pequenos efeitos diferenciais negativos. Implicações para pesquisa, política e prática são discutidas.

Palavras-chave: Avaliação baseada em desempenho; realização acadêmica; *accountability*; política educacional

Introduction

Since the passage of the No Child Left Behind (NCLB) Act in 2001, states have been required to annually report K-12 students' proficiency in English language arts (ELA) and math in grades 3-8 and once in high school using state annual standardized achievement tests. However, NCLB's high-stake sanctions for schools failing to meet adequate yearly progress towards 100% proficiency in ELA and math by 2013-14, as required by the law, created many unintended consequences. Teachers tended to 1) narrow the curriculum to focus on just those subjects tested on state assessments, 2) present content in fragmented test-related pieces, 3) increase the use of teacher-centered instructional practices, and 4) sometimes even try to "game the system" by focusing on students near the proficiency cut off or changing student answers (Au, 2007; Booher-Jennings, 2005; Diamond & Spillane, 2004; Hamilton et al., 2007; McMurrer, 2007; Stecher et al., 2008).

The Every Student Succeeds Act (ESSA) replaced NCLB in 2015. Although ESSA continues requirements for state annual achievement testing in grades 3-8 and once in high school in ELA and math, the new law eliminates the school accountability sanctions. ESSA also provides an avenue for up to seven states (or groups of them) to establish and operate innovative assessment systems, including for use in statewide accountability systems. The term "innovative" is used to differentiate state assessment systems that do not rely solely on statewide annual achievement tests to determine student proficiency each year in the required grades and subjects. Instead, states can experiment with other types of assessment such as performance-based, competency-based, and interim assessments to meet the federal requirements. This provision in the law, known as the Innovative Assessment Demonstration Authority, is in response to the perceived negative curricular and instructional consequences of state standardized testing, as well as the widespread belief that "what gets measured is what gets taught." 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).

And yet there is little empirical evidence on the efficacy of performance-based assessment to improving student achievement outcomes in a school accountability context. Prior to the passage of NCLB, which contributed to the demise of some states' performance assessment programs (Stecher, 2010), Shepard and colleagues (1995) stated that 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 and not from any research-based evidence.

Therefore, the purpose of this descriptive study was to examine the effects on student achievement outcomes of a proof of concept pilot program that utilizes performance-based assessments to make determinations of student proficiency in a school accountability context. New Hampshire's Performance Assessment of Competency Education (PACE) pilot program was officially approved by the U.S. Department of Education in March 2015 and currently operates under a first-in-the-nation waiver from federal statutory requirements related to state annual achievement testing (New Hampshire Department of Education [NHDOE], 2015). The Innovative Assessment Demonstration Authority was inspired in large measure by NH's PACE pilot (Klein, 2016). As of the writing of this article, PACE is now in its fifth year of implementation (2014-15 to 2018-19 school years) and NH's application under the Demonstration Authority was approved in fall 2018. This study examines student achievement outcomes after the first three years. The only other research on the PACE pilot was a formative evaluation that did not examine outcomes, but instead provided feedback to the PACE Leadership team to monitor and adjust implementation (Becker et al., 2017).

PACE is considered an "innovative" assessment system because it does not rely on state annual standardized achievement tests to make determinations of student proficiency in the required grades and subjects. Annual determinations of student proficiency in PACE districts are based on local summative assessment data from grade books (including common and local performance-based assessments) alongside teacher judgments about student achievement levels, except in those grades and subject areas where the state achievement test is administered (Marion & Leather, 2015). There is a state-level achievement test administered once per grade span that acts as an external audit on the system.

The present study is de-limited to Grade 8 and Grade 11 students because of the way this specific innovative assessment system is designed. Districts self-select into the PACE pilot, which means there are two state assessment systems currently operating within NH that are used to produce annual determinations of student proficiency as required by federal law—the statewide assessment system that uses an annual standardized achievement test and the PACE assessment system that uses local assessment information and teacher judgments. The only assessment in common between the two systems that can be used to examine effects is the annual standardized achievement test taken by students in the PACE pilot once per grade span: grade 3 ELA, grade 4 math, and grades 8 and 11 ELA and math.

The state standardized achievement test is an adequate and appropriate measure of the early effects of an innovative assessment system such as PACE because one key criterion for the PACE pilot is "no harm" on the state achievement test. System designers believe the state achievement test, though very different from the use of performance-based assessments to determine student proficiency, provides evidence that students in PACE schools/districts are provided an equitable opportunity to learn the content standards (NHDOE, 2014b). Prior research on mastery learning initiatives in the 1970s and 1980s engendered debates about the best outcome measure to use to examine effects of programs intended to deepen students' exposure to content such as the use of performance assessments. Some researchers at that time, particularly Slavin (1987), argued that standardized measures show that the breadth of content coverage is not sacrificed in the quest for content depth and should be preferred over researcher-created or locally-created outcomes measures.

Additionally, Grades 8 and 11 were chosen because there is no prior achievement data available for grade 3 ELA or grade 4 math, which is why those grades were not examined. 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). There were three research questions:

(1) What is the average effect of the PACE pilot on Grade 8 and 11 student achievement in mathematics and English language arts?
(2) To what extent do effects vary for certain subgroups of students?
(3) To what extent does the number of years a district has implemented the PACE pilot affect student achievement outcomes?

Review of Related Research

Performance 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 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, Baker, & Dunbar, 1991).

Over time researchers have collected evidence on the benefits, limitations, and lessons learned from the implementation of state-adopted performance assessment systems in 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). Much of the research on performance assessment programs specific to student achievement outcomes focuses on relations between teacher-reported changes in instructional practices and student achievement outcomes (Lane, Parke, & Stone, 2002; Parke, Lane, & Stone, 2006; Stecher, Barron, Kaganoff, & Goodwin, 1998; Stecher, Barron, Chun, & Ross, 2000; Stecher & Chun, 2001; Stone & Lane, 2003). However, there is some empirical evidence 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 prior studies that examined the effects of performance assessments 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 multiple subject areas, including: math, reading and writing (Stone & Lane, 2003).

There has been one formative evaluation of the PACE pilot to date (Becker et al., 2017). The primary goal of this formative evaluation was to provide feedback and information to the PACE Leadership team on the implementation process for continuous improvement purposes. The evaluation investigated four interim goals/claims: 1) stakeholders are committed to PACE; 2) performance assessments are based on sound test design principles; 3) performance assessments are successfully implemented; and 4) scores are accurate and reliable. Evaluators used qualitative and quantitative data collected from a teacher survey, site visits, classroom observations, and focus group interviews with students, parents, teachers, and administrators during the 2016-17 school year. Overall, the evaluation found significant buy-in and collaboration amongst all stakeholders involved and consistent evidence of the PACE teachers' assessment literacy. Most of the evidence also supported the claim that implementing performance tasks had a positive impact on instruction. Teachers expressed how implementing performance tasks increased the cognitive rigor of their teaching and provided real-time feedback to adjust and monitor instruction to better meet the needs of students. Students noted how PACE requires more real-world application and the ability to demonstrate their knowledge and skills in multiple ways. Parents noted that PACE common performance tasks encouraged a deeper level of understanding than a traditional multiple choice test and they believed their student retained their learning longer. Ten recommendations were made in the final report to improve the quality of the implementation process and inform decision-making, including the call for research that externally verifies the impacts of PACE on teaching and learning (Becker et al., 2017, p. 33).

Description of Intervention

In the fall of 2014, the NHDOE applied for a 2-year waiver (2014-2015 and 2015-2016 school years) from NCLB's 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 first-in-the-nation waiver in March 2015, allowing self-selected NH school districts to base annual determinations of student proficiency in ELA and math in grades 3-8 and once in high school on a combination of local, common, and state-level assessments (NHDOE, 2014a). The pilot was granted additional one-year waivers for the 2016-17 and 2017-18 school years. The NHDOE applied and received permission under the Innovative Assessment Demonstration Authority in fall 2018 to continue and scale the pilot statewide from the 2018-19 school year to the 2025-26 school year.

Table 1 provides an overview of the PACE assessment system. Local assessments include all summative assessments (including performance assessments) given within districts to assess student progress towards proficiency (or 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 (Evans & Lyons, 2017). In the PACE pilot, state-level achievement testing occurs once

Table 1
Overview of the New Hampshire PACE Assessment and Accountability System

Grade	ELA	Math
3	Statewide achievement test (SBAC)	Local and common performance assessments (PACE)
4	Local and common performance assessments (PACE)	Statewide achievement test (SBAC)
5	Local and common performance assessments (PACE)	Local and common performance assessments (PACE)
6	Local and common performance assessments (PACE)	Local and common performance assessments (PACE)
7	Local and common performance assessments (PACE)	Local and common performance assessments (PACE)
8	Statewide achievement test (SBAC)	Statewide achievement test (SBAC)
11	Statewide achievement test (SAT)	Statewide achievement test (SAT)

Note. Table adapted from NHDOE, 2016b.

per grade span: grade 3 ELA, grade 4 math, and grades 8 and 11 ELA and math. 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 (NHDOE, 2016b).

The process for school districts 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).

The PACE pilot continues to scale each year. Currently, 14 out of 84 SAUs (School Administrative Units) in the state implement the PACE pilot for accountability purposes, but there is a rolling cohort structure: four SAUs joined in 2014-15, four more joined in 2015-16, one joined in 2016-17, and five joined in 2017-18. This allows student dosage effects to be examined by the number of years the district has implemented the PACE pilot.

PACE Theory of Action

According to the NHDOE (2016b), 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 1 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 broadly illustrate 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. Although this description of the PACE theory of action could describe a lot of state accountability efforts, the PACE system differs from previous assessment and accountability efforts in NH due to three design features.

The first design feature is that local education leaders are explicitly involved in designing and implementing their own accountability system, which is not true outside of the PACE system. This is intended to foster positive organizational learning and change by supporting the internal motivation of educators. System designers believe this contrasts with 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 access to capacity building training and resources to support their development of key capacities related to designing and implementing the system. These supports and resources are not available to non-PACE districts in the state. This means the NH DOE and its technical partners provide professional development, training, and support to local districts in the technical, policy, and practical issues related to the system design and implementation—particularly the design and consistent scoring of common performance assessments and other assessment literacy topics.

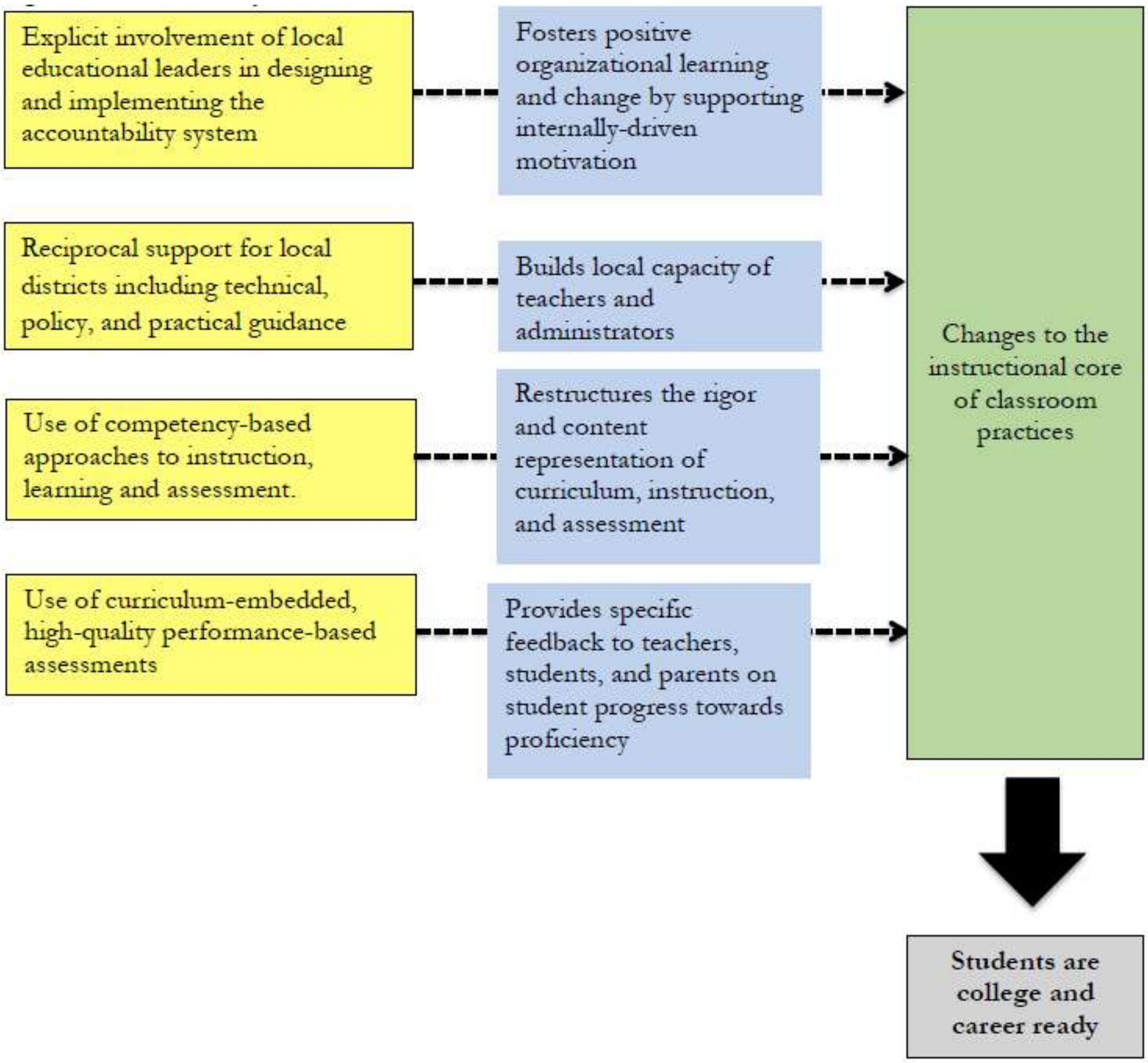
Explicit involvement of local educational leaders in designing and implementing the accountability system
Fosters positive organizational learning and change by supporting internally-driven motivation
Reciprocal support for local districts including technical, policy, and practical guidance
Builds local capacity of teachers and administrators
Use of competency-based approaches to instruction, learning and assessment.
Restructures the rigor and content representation of curriculum, instruction, and assessment
Use of curriculum-embedded, high-quality performance-based assessments
Provides specific feedback to teachers, students, and parents on student progress towards proficiency
Changes to the instructional core of classroom practices
Students are college and career ready

Figure 1. PACE Theory of Action

The third design feature is the use of locally designed and curriculum-embedded performance assessments throughout the year. According to PACE system designers, performance 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 performance tasks 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, which is designed collaboratively by all the participating PACE districts, is intended to serve as an exemplar for teachers as they design local performance assessments and rubrics. As more PACE Common Tasks are designed, there is a bank of performance tasks and rubrics with anchor papers at different levels of performance available to PACE districts.

The ultimate goal of PACE is that student achievement outcomes improve and that all students are college or career ready upon graduation from high school (NHDOE, 2016b). It is important to note that the NHDOE's *a priori* expectation for student achievement outcomes resulting from the PACE pilot over the first few years was "no harm" on the state achievement test. This criterion was defined as "the performance of students in PACE districts does not decline compared to predicted scores on Smarter Balanced [and SATs]" (NHDOE, 2015, p. 5). As stated earlier, the criterion of "no harm" on Smarter Balanced and SATs provides evidence that students in PACE schools/district are provided an equitable opportunity to learn the content standards because both state achievement tests are aligned with the State's grade and subject level content standards and model competencies. "Equitable opportunity to learn the content standards" is a term often used in the standards-based reform and test-based accountability movements because of equity concerns. The argument that civil rights advocates often make is that standardized achievement tests shed light on achievement gaps because students cannot perform well on a standardized achievement test aligned to the state's content standards if they have not been taught the content standards. Therefore, equitable opportunity to learn the content standards is an inference made based on student achievement test results and is a key criterion used to examine the outcomes of state and federal educational assessment and accountability reform policies.

Method

Data Description

The data for this study come from student-level data sets provided by the NHDOE. The NHDOE provided scale score results for all students in NH (grades 3-11) who completed a state achievement test in math or ELA for the 2011-12 (Baseline), 2012-13 (Baseline), 2013-14 (Baseline), 2014-15 (Year 1 Outcome), 2015-16 (Year 2 Outcome), and 2016-17 (Year 3 Outcome) school years, along with school and district ID information. Students were removed from the data set if they were retained in Gradex 8 or 11. NH switched state achievement tests in spring 2015 so all of the baseline data sets have scale score results from the New England Common Assessment Program (NECAP) and all of the outcome data sets have results from the Smarter Balanced Assessment Consortium (SBAC) or SATs (which didn't start until spring 2016). The data sets also included student background and demographic characteristics such as disability status, gender, free-and reduced price lunch (FRL) status, limited English proficiency (LEP) status, and race/ethnicity. This analysis controls for these student-level characteristics to improve effect estimate precision.

Sample Selection Process

The outcome data sets were merged with the baseline data sets so that each student in Grades 8 and 11 had a matched prior achievement scale score in math or ELA from the most recent year of prior testing available for the sample. Creating separate matched data sets for each outcome maximized the sample size for each outcome analysis, as patterns of missing data varied across outcome. Students were removed by grade and subject area from the analytic sample if they attended special education private schools, did not have any baseline prior achievement test results, and had missing student background/demographic information. This resulted in trimming of the outcome data sets by approximately 10% (Grade 8 ELA=10.35%; Grade 8 Math=10.35%; Grade 11 ELA=9.66%; Grade 11 Math=9.92%). Table 2 contains the number of districts, schools, and students in the treatment (PACE) and comparison group (non-PACE) each school year and number of treatment years by grade and subject in the unweighted analytic sample. PACE status is allowed to differ by outcome year for districts/schools as they enter PACE in different years. Because of the small treatment (PACE) sample size (4-9 districts), there is limited power to detect differences in the population from which the sample is drawn unless they are very large differences.

Table 2
Number of Districts, Schools, and Students in the PACE (Treatment) and Non-PACE (Comparison Group) by School Year or Number of Treatment Years in the Unweighted Analytic Samples

	2014-15	2015-16	2016-17	Treat1	Treat2	Treat3	*N*_Students Total
PACE (Treatment)							
N_Districts	4	8	9	9	8	4	
N_Schools	4	8	9	9	8	4	
Gr 8 Math	487	750	855	848	805	439	2,092
Gr 8 ELA	482	751	864	842	806	449	2,097
Gr 11 Math	637	975	1,009	1,034	939	648	2,621
Gr 11 ELA	658	970	1,008	1,050	937	649	2,636
Non-PACE (Comparison Group)							
N_Districts	103	99	98	NA	NA	NA	
N_Schools (Gr 8/Gr 11)	141/88	137/84	136/83	NA	NA	NA	
Gr 8 Math	11,906	11,291	10,899	NA	NA	NA	34,096
Gr 8 ELA	11,887	11,303	10,903	NA	NA	NA	34,093
Gr 11 Math	9,421	10,101	10,342	NA	NA	NA	29,864
Gr 11 ELA	9,519	10,090	10,332	NA	NA	NA	29,941

Note. NA=not applicable; Treat1=one year of treatment; Treat2=two years of treatment, and Treat3=three years of treatment

Baseline Characteristics of the Unweighted Analytic Samples

Selection is at the district-level because districts made the decision to implement the PACE pilot. As a result, it is important to establish baseline equivalence for the PACE and non-PACE comparison groups in the analytic sample using characteristics at that same level (i.e., district-level; Institute of Education Sciences, 2014). 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 baseline data files by year to capture pre-treatment differences in districts for those students in the analytic sample. These six district-level characteristics are plausibly related to outcome and include: the percent of (1) IEP students in the district, (2) FRL students in the district, (3) LEP students in the district, (4) non-White students in the district, (5) students proficient or above in math, (6) students proficient or above in ELA. Since these variables were aggregated from the student-level baseline data sets they only include students in tested grades (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 or Grade 11 students in one district have the same district-level percent by characteristic and year. A weighted average for each of the district characteristics was then computed by treatment status (non-PACE=comparison group; PACE=Treatment group) using all the students in the analytic sample by grade and subject area.

Tables 3-4 provide the baseline (unweighted) district and student characteristics by subject area and treatment status for Grades 8 and 11, respectively. Standardized mean differences (SMDs) between the groups on each district and student characteristic were calculated and *t*-tests were used to examine whether the mean differences were statistically different between the groups. According to the *What Works Clearinghouse* Group Design Standards (Institute of Education Sciences, 2014), SMDs in absolute value 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).

Overall, there are a few notable differences when comparing the treatment (PACE) and comparison (non-PACE) groups as evidenced by SMDs greater than 0.25. First, across all grades and subject areas there tends to be higher average district percentages of IEP, FRL, and LEP students in the PACE group. Second, PACE districts tend to have lower percentages of students who are proficient or above in math and ELA than the comparison group. 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 all of the observed district characteristics do not satisfy baseline equivalence (SMD<0.05), inverse propensity score weighting was employed to attempt to balance the two groups on the observable district characteristics prior to outcome analyses (Guo & Fraser, 2015; Murnane & Willett, 2011).

Table 3
Grade 8 Math and ELA Baseline Characteristics of the Unweighted and Inverse Propensity Score Weighted Analytic Samples on District and Student Characteristics by Treatment Status

	Grade 8											
	Math						ELA					
	Non-PACE Comparison		PACE Treatment				Non-PACE Comparison		PACE Treatment			
	M	(SD)	M	(SD)	SMD	Sign.	M	(SD)	M	(SD)	SMD	Sign.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
District Characteristics Unweighted Analytic Samples												
N_students	34,096		2,092				34,093		2,097			
%IEP	0.15	(.03)	0.17	(.03)	-0.64	***	0.15	(.03)	0.17	(.03)	-0.64	***
%FRL	0.26	(.16)	0.34	(.12)	-0.62	***	0.26	(.16)	0.34	(.12)	-0.62	***
%LEP	0.02	(.02)	0.02	(.02)	-0.05	***	0.02	(.02)	0.02	(.02)	-0.06	***
%Non-White	0.11	(.09)	0.10	(.04)	0.16	***	0.11	(.09)	0.10	(.04)	0.15	***
%Math Prof	0.67	(.10)	0.62	(.05)	0.64	***	0.67	(.10)	0.62	(.05)	0.64	***
%ELA Prof	0.78	(.08)	0.71	(.05)	0.99	***	0.78	(.08)	0.71	(.05)	1.00	***
District Characteristics Inverse Propensity Score Weighted Analytic Samples												
Wtd. N_students	36,374		31,672				36,379		31,815			
%IEP	0.15	(.03)	0.14	(.03)	0.31	***	0.15	(.03)	0.14	(.03)	0.32	***
%FRL	0.27	(.16)	0.29	(.10)	-0.21	***	0.27	(.16)	0.29	(.10)	-0.21	***
%LEP	0.02	(.03)	0.02	(.02)	-0.13	***	0.02	(.03)	0.02	(.02)	-0.14	***
%Non-White	0.11	(.09)	0.09	(.05)	0.23	***	0.11	(.09)	0.09	(.05)	0.23	***
%Math Prof	0.66	(.11)	0.66	(.07)	0.03	**	0.66	(.11)	0.66	(.07)	0.03	*
%ELA Prof	0.77	(.09)	0.77	(.06)	0.12	***	0.77	(.09)	0.77	(.06)	0.12	***

Table 3 (Cont'd.)

Grade 8 Math and ELA Baseline Characteristics of the Unweighted and Inverse Propensity Score Weighted Analytic Samples on District and Student Characteristics by Treatment Status

	Grade 8											
	Math						ELA					
	Non-PACE Comparison		PACE Treatment				Non-PACE Comparison		PACE Treatment			
	M (1)	(SD) (2)	M (3)	(SD) (4)	SMD (5)	Sign. (6)	M (7)	(SD) (8)	M (9)	(SD) (10)	SMD (11)	Sign. (12)
Student Characteristics Unweighted Analytic Samples												
N_students	34,096		2,092				34,093		2,097			
%Male	0.52	(.50)	0.49	(.50)	0.12	*	0.52	(.50)	0.49	(.50)	0.06	*
%Non-White	0.11	(.31)	0.09	(.29)	0.07	**	0.11	(.31)	0.09	(.28)	0.07	**
%IEP	0.13	(.34)	0.15	(.35)	-0.06		0.13	(.34)	0.15	(.35)	-0.06	
%FRL	0.24	(.43)	0.33	(.47)	-0.20	***	0.24	(.43)	0.33	(.47)	-0.20	***
%LEP	0.01	(.09)	0.01	(.09)	0.00		0.01	(.09)	0.01	(.09)	0.00	
Student Characteristics Inverse Propensity Score Weighted Samples												
Wtd. N_students	36,374		31,672				36,379		31,815			
%Male	0.52	(.50)	0.49	(.50)	0.06	***	0.52	(.50)	0.49	(.50)	0.06	***
%Non-White	0.11	(.31)	0.08	(.27)	0.10	***	0.11	(.31)	0.08	(.27)	0.10	***
%IEP	0.14	(.34)	0.11	(.32)	0.09	***	0.14	(.34)	0.11	(.32)	0.09	***
%FRL	0.24	(.43)	0.30	(.46)	-0.14	***	0.24	(.43)	0.30	(.46)	-0.14	***
%LEP	0.01	(.09)	0.01	(.09)	0.00		0.01	(.09)	0.01	(.09)	0.00	

Note. All district characteristics were aggregated from baseline data sets and merged into outcome data sets 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; %Non-White=percent of students classified as American Indian/Alaskan Native, Asian, Black, Hispanic, Native Hawaiian/Pacific Islander, and Two or more races; %Math Prof=percent of students proficient or above in math in the district; %ELA Prof=percent of student proficient or above in ELA in the district; M=mean; SD=standard deviation; SMD=standardized mean difference using pooled standard deviations; Sign.=p-value.

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

Table 4
Grade 11 Math and ELA Baseline Characteristics of the Unweighted and Inverse Propensity Score Weighted Analytic Samples on District and Student Characteristics by Treatment Status

	Grade 11											
	Math						ELA					
	Comparison		Treatment				Comparison		Treatment			
	M	(SD)	M	(SD)	SMD	Sign.	M (7)	(SD)	M (9)	(SD)	SMD	Sign.
	(1)	(2)	(3)	(4)	(5)	(6)		(8)		(10)	(11)	(12)
District Unweighted Analytic Samples												
N_students	29,864		2,621				29,941		2,636			
%IEP	0.18	(.10)	0.21	(.11)	-0.30	***	0.18	(.10)	0.21	(.11)	-0.30	***
%FRL	0.17	(.17)	0.19	(.18)	-0.10	***	0.17	(.17)	0.19	(.18)	-0.11	***
%LEP	0.06	(.07)	0.07	(.07)	-0.10	***	0.06	(.07)	0.07	(.07)	-0.10	***
%Non-White	0.10	(.09)	0.09	(.05)	0.25	***	0.10	(.09)	0.09	(.05)	0.26	***
%Math Prof	0.62	(.13)	0.58	(.09)	0.37	***	0.62	(.13)	0.58	(.09)	0.38	***
%ELA Prof	0.79	(.08)	0.76	(.06)	0.47	***	0.79	(.08)	0.76	(.06)	0.47	***
District Inverse Propensity Score Weighted Analytic Samples												
Wtd. N_students	32,690		28,588				32,790		28,619			
%IEP	0.18	(.11)	0.20	(.11)	-0.15	***	0.18	(.11)	0.20	(.11)	-0.15	***
%FRL	0.17	(.17)	0.17	(.16)	0.00		0.17	(.17)	0.17	(.17)	0.00	
%LEP	0.06	(.07)	0.07	(.07)	-0.11	***	0.06	(.07)	0.07	(.07)	-0.12	***
%Non-White	0.10	(.09)	0.09	(.05)	0.13	***	0.10	(.09)	0.09	(.05)	0.13	***
%Math Prof	0.62	(.14)	0.58	(.10)	0.27	***	0.62	(.14)	0.58	(.10)	0.27	***
%ELA Prof	0.79	(.09)	0.77	(.06)	0.16	***	0.79	(.09)	0.77	(.06)	0.16	***

Table 4 (Cont'd.)

Grade 11 Math and ELA Baseline Characteristics of the Unweighted and Inverse Propensity Score Weighted Analytic Samples on District and Student Characteristics by Treatment Status

	Grade 11											
	Math						ELA					
	Comparison		Treatment				Comparison		Treatment			
	M	(SD)	M	(SD)	SMD	Sign.	M (7)	(SD)	M (9)	(SD)	SMD	Sign.
	(1)	(2)	(3)	(4)	(5)	(6)		(8)		(10)	(11)	(12)
Student Unweighted Analytic Samples												
N_students	29,864		2,621				29,941		2,636			
%Male	0.51	(.50)	0.51	(.50)	0.00		0.51	(.50)	0.51	(.50)	0.00	
%Non-White	0.09	(.29)	0.08	(.27)	0.04	**	0.09	(.29)	0.07	(.26)	0.07	**
%IEP	0.12	(.32)	0.12	(.33)	0.00		0.12	(.32)	0.12	(.33)	0.00	
%FRL	0.19	(.39)	0.20	(.40)	-0.03		0.19	(.39)	0.20	(.40)	-0.03	
%LEP	0.00	(.07)	0.01	(.09)	-0.13		0.00	(.06)	0.01	(.07)	-0.15	
Student Inverse Propensity Score Weighted Analytic Samples												
Wtd. N_students	32,690		28,588				32,790		28,619			
%Male	0.51	(.50)	0.52	(.50)	-0.02		0.51	(.50)	0.51	(.50)	0.00	
%Non-White	0.09	(.29)	0.08	(.28)	0.04	**	0.09	(.29)	0.08	(.27)	0.04	***
%IEP	0.12	(.32)	0.11	(.32)	0.03	*	0.12	(.32)	0.11	(.32)	0.03	
%FRL	0.19	(.39)	0.18	(.38)	0.03	**	0.19	(.39)	0.17	(.38)	0.05	***
%LEP	0.00	(.07)	0.01	(.09)	-0.12	***	0.00	(.06)	0.01	(.08)	-0.14	**

Note. All district characteristics were aggregated from baseline data sets and merged into outcome data sets 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; %Non-White=percent of students classified as American Indian/Alaskan Native, Asian, Black, Hispanic, Native Hawaiian/Pacific Islander, and Two or more races; %Math Prof=percent of students proficient or above in math in the district; %ELA Prof=percent of student proficient or above in ELA in the district; M=mean; SD=standard deviation; SMD=standardized mean difference using pooled standard deviations; Sign.=p-value.

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

Propensity Score Estimation

Propensity score methods allow a researcher to potentially 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).

Propensity score model. 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): $e(x_i) = \mathrm{pr}(W_i = 1|X_i = x_i)$. To estimate propensity scores, a binary logistic regression model was specified that included six observed district characteristics plausibly related to outcome discussed above. Because district 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.

Inverse propensity score weighting. Inverse propensity score weighting (IPSW) 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). It does so by weighting down treatment students with large estimated probabilities and comparison students with small estimated probabilities and vice versa.
Average treatment effect (ATE) weights were calculated using the following equation:

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

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

IPSW was chosen over other propensity score methods because each student in a district had the same estimated propensity score, which would have resulted in significant trimming of the sample if certain propensity score matching methods were utilized, whereas, IPSW maximized the sample size for each outcome analysis.

Baseline Chracteristics of the IPSW Analytic Sample

Tables 3-4 provide baseline characteristics of the Grade 8 and Grade 11 IPSW math and ELA analytic samples on observed district and student characteristics by treatment status. There are differences between the two groups on almost all district characteristics even after IPSW (SMD>0.05), although the weighting does create more equivalent groups and therefore will be used as probability weights in subsequent statistical analyses. When examining student-level characteristics prior to outcome analyses, the PACE and non-PACE comparison group were fairly equivalent, but not equivalent. Other propensity score matching techniques were examined such as nearest neighbor matching within a caliper, but other approaches did not eliminate the equivalence issues and reduced the sample size. Including additional variables in the propensity score model was not possible given the limited information available from the NHDOE.

These differences between the PACE treatment and non-PACE comparison group even after IPSW suggest that any findings resulting from subsequent multivariate analyses should be considered descriptive rather than causal. Students attend 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. As a result of the two groups (PACE vs. non-PACE) not being equivalent prior to outcome analyses, this research's original aim at producing an average treatment effect (ATE) estimate was abandoned and instead descriptive results that estimate differences in performance between the two non-equivalent groups are provided. Descriptive results illuminate differences in performance between the PACE group and non-PACE comparison group, but do not mean that PACE treatment *causes* either better performance or worse performance on the state achievement test. Findings are instead correlational.

Measures

Three different types of variables are included in the analyses: (1) student-level outcome variables, (2) student-level control variables, and (3) school-level treatment and control variables. Students are nested within schools that are also nested within school districts in this study; however, all PACE districts have only one middle or high school so school-level effects and district-level effects are confounded. School-level treatment and controls were chosen to model instead of district-level variables 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 (Hanushek, Kain, Markman, & Rivkin, 2001).

The student-level outcomes were Smarter Balanced (SBAC) math and ELA scale scores in Grade 8 and SAT math and ELA scale scores in Grade 11, all of which were converted to *z*-scores by year of outcome. The student-level control variables included grand-mean centered standardized prior achievement in math or ELA (depending on outcome variable). Table 5 contains descriptive statistics on the standardized student-level outcome and prior achievement variables by year and treatment group. The spread and variability of the standardized variables were similar across the two groups with the outcome variable being more constricted in range than the prior achievement variable.

Student-level control variables also included a series of dummy variables: free-or-reduced price lunch status (FRL=1), disability status (IEP=1), and gender (male=1). Centering allows the intercept to be interpretable. Limited English proficiency (LEP) status and race/ethnicity variables were also examined but ultimately not included because there was only 1% LEP and 10% non-White students in the analytic sample, which mirrors state demographics, but can lead to misleading findings. For example, low percentages of non-White students is an issue in this sample because lumping all non-White students into one dummy variable can produce misleading results. This is because in some New Hampshire communities the largest percentages of non-White students self-identify as Asian and some Asian subgroups tend to score higher on achievement tests than other non-White subgroups. This leads to results where non-White students appear to perform better, on average, than White students, but it is the self-identified Asian group driving the positive effects.

Table 5

Descriptive Characteristics of the Standardized Outcome and Prior Achievement Variables by Year and Treatment Status

			Gr 8 Math			Gr 11 Math			Gr 8 ELA			Gr 11 ELA		
Year ID	Treatment Group	Standardized Variable	N	Min	Max	N	Min	Max	N	Min	Max	N	Min	Max
1415	Non-PACE	Outcome	11906	-2.95	2.28	9421	-2.56	2.28	11887	-3.36	2.14	9519	-2.92	1.82
		Prior Achvmt	11906	-4.07	3.11	9421	-4.80	3.66	11887	-3.74	2.41	9519	-4.29	2.29
	PACE	Outcome	487	-2.95	2.28	637	-2.56	2.28	482	-3.22	2.08	658	-2.92	1.82
		Prior Achvmt	487	-4.07	2.30	637	-4.80	2.60	482	-3.74	2.41	658	-4.29	2.29
1516	Non-PACE	Outcome	11291	-3.01	2.16	10101	-3.09	2.88	11303	-3.42	1.99	10090	-3.37	2.82
		Prior Achvmt	11291	-3.80	2.85	10101	-4.66	3.70	11303	-4.02	2.55	10090	-4.23	2.29
	PACE	Outcome	750	-3.01	2.16	975	-3.09	2.58	751	-3.36	1.99	970	-2.64	2.72
		Prior Achvmt	750	-3.80	2.27	975	-4.66	3.07	751	-3.28	2.55	970	-4.23	2.29
1617	Non-PACE	Outcome	10899	-2.90	2.17	10342	-3.09	2.77	10903	-3.21	2.05	10332	-3.23	2.86
		Prior Achvmt	10899	-4.27	3.10	10342	-4.45	3.56	10903	-4.43	2.86	10332	-4.02	2.30
	PACE	Outcome	855	-2.90	2.17	1009	-3.09	2.77	864	-2.92	2.05	1008	-2.51	2.45
		Prior Achvmt	855	-4.00	3.10	1009	-4.45	3.56	864	-4.43	2.86	1008	-4.02	2.30

Note. The variables are standardized so they have a mean of 0 and standard deviation of 1.

The school-level control variables were all aggregated from baseline student covariates by school and included school mean prior ELA or math achievement (depending on subject area of outcome). Other school-level controls such as the percentage of FRL, IEP, LEP, and non-White students in the school were also examined, but were ultimately removed because of either low mean percentages of students within schools or poor model fit. All school-level control variables were grand-mean centered to aid interpretation of the intercept. Two dummy variables for outcome year were also included (Year1516 and Year1617—the 2014-15 school year being the reference category).

Differences in the performance of the two non-equivalent groups (PACE vs. non-PACE) were modeled in two independent ways depending on the model. First, a binary indicator for treatment (pace=1) indicated whether a school was implementing the PACE pilot in a particular year. Second, differences in performance were modeled using three dummy variables (treat1, treat2, or treat3) that vary by year indicating whether a school was in its first, second, or third year of implementation in a given year. This allowed for an examination of dosage effects. Only students who remained in their current school during all years of PACE implementation are included in the treat2 and treat3 dummy variables.

Analytic Approach

In order to address the first and second research questions, the analyses began with estimating the effect of the PACE pilot on cross-sectional Grades 8 and 11 student cohorts and subgroups. Since students are nested within schools, multi-level modeling (Raudenbush & Bryk, 2002) was used to estimate the average effect of the NH PACE pilot on the student-level outcome variables. A three-level model (students nested within schools and schools nested within districts) is not possible given that only a couple districts have more than one middle or high school. The intraclass correlation coefficient suggests that about 8% of the variance in Grades 8 and 11 math and ELA achievement in this sample is between schools and the other 92% is within schools. 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.

Grade by subject area achievement was modeled separately and student- and school-level predictors and controls were used to account for differences in students' baseline test scores and demographic/ background characteristics. Equation 1 is shown with standard notation for a two-level composite model:

$$\begin{aligned} &Y_{ij} \\ &= \beta_0 + \beta_1 \boldsymbol{X}_{ij} + \beta_2 \boldsymbol{S}_j + \beta_3(\mathrm{PACE}_j) + \beta_4(\mathrm{Year1516}_j) + \beta_5(\mathrm{Year1617}_j) \\ &+ \beta_6(\mathrm{PACE}_j * \mathrm{Year1516}_j) + \beta_7(\mathrm{PACE}_j * \mathrm{Year1617}_{ij}) + \beta_8(\mathrm{PACE}_j * \boldsymbol{X}_{ij}) + u_{0j} + r_{ij} \end{aligned} \tag{1}$$

where grade by subject area achievement of student *i* in school *j* (Y_{ij}) is a function of a vector of that student's observable characteristics such as prior achievement, gender, disability status, free-and-reduced price lunch status $(\mathbf{X}_{ij})$, school characteristics such as school prior achievement in ELA or math $(\mathbf{S}_j)$, binary treatment variable (PACE_j), Year ID $(\mathrm{Year1516}_j$ or $\mathrm{Year1617}_j)$, interactions between treatment and Year ID, interactions between treatment 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}). The only random effect in this model is the intercept, although all student-level variables were examined as random effects in the model building process.

For the first research question, the parameter estimates of interest are those associated with treatment and the interactions between treatment and Year ID. For the second research question, the parameter estimates of interest are the cross-level interactions between treatment and student-level characteristics. This provides insight into whether certain subgroups of 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. Given the relatively small PACE sample size overall and the fact that the interaction analyses cut the data into even smaller slices, the interaction results should be considered somewhat exploratory. However, the interaction results are worth noting because of the similarity in some of the interaction findings across years.

In order to address the third research question, "To what extent does the number of years a district has implemented the PACE pilot affect student achievement outcomes?" analyses focused on estimating dosage effects by year. Equation 2 specified the two-level composite model:

$$\begin{aligned} Y_{ij} \\ &= \beta_0 + \beta_1 \boldsymbol{X}_{ij} + \beta_2 \boldsymbol{S}_j + \beta_3(\text{treat1}_j) + \beta_4(\text{treat2}_j) + \beta_5(\text{treat3}_j) + \beta_6(\text{Year1516}_j) \\ &+ \beta_7(\text{Year1617}_j) + \beta_8(\text{treat1}_j * \text{Year1516}_j) + \beta_9(\text{treat1}_j * \text{Year1617}_j) \\ &+ \beta_{10}(\text{treat2}_j * \text{Year1617}_j) + u_{0j} + r_{ij} \end{aligned} \tag{2}$$

where grade by subject area achievement of student *i* in school *j* (Y_{ij}) is a function of a vector of that student's observable characteristics, school characteristics, treatment indicating either one, two, or three years of dosage, Year ID, interactions between treatment dosage and Year ID, the random effect of the intercept, and a residual term that captures the random noise that may occur at the student-level. For the third research question, the parameter estimates of interest are those associated with treatment dosage and interactions between treatment dosage and Year ID. Subgroup analysis was not modeled or examined in Equation 2 because the number of students in some subgroups for a particular treatment dosage in a given year was less than 10 students.

Results

RQ#1: Average Effect

In the analyses, multilevel models were used to estimate effects using the cross-sectional data. The regression estimates associated with treatment (PACE) are mean differences in SD units between treatment and comparison groups. Positive estimates indicated a positive effect. The parameter estimates and variance components from Model I (Equation 1) are reported in Table 6. The ATE inverse propensity score weight was used as a regression weight in all models.

Effect estimates resulting from Model I were positive, which suggests that there is an overall positive effect of treatment in Grades 8 and 11 math and ELA over the first three years of implementation. There does not appear to be an implementation dip over the first three years, which can be ascertained from the parameter estimates associated with the interactions between treatment and Year ID. The estimates obtained from Grade 8 were on average around 14% of a standard deviation in math (*d*=~0.14), but smaller in ELA (*d*=~0.06) based on the average New Hampshire student. The reverse is true in Grade 11 where effect estimates for the average New Hampshire student were very small in math (*d*=~0.03), but larger in ELA (*d*=~0.09).

Table 6

Model I Parameter Estimates and Variance Components from Multilevel Models Showing the Conditional Effects of PACE on Grade 8 and 11 Math and ELA Achievement Using Inverse Propensity Score Weights

	Grade 8 Math			Grade 8 ELA			Grade 11 Math			Grade 11 ELA		
	Estimate		SE	Estimate		SE	Estimate		SE	Estimate		SE
	(1)		(2)	(3)		(4)	(5)		(6)	(7)		(8)
Intercept	0.14	***	(.02)	0.22	***	(.02)	0.00		(.01)	0.04	*	(.02)
Prior Achvmt	0.72	***	(.0)	0.56	***	(.01)	0.72	***	(.01)	0.62	***	(.01)
IEP	-0.20	***	(.01)	-0.38	***	(.02)	-0.10	***	(.02)	-0.31	***	(.02)
Male	-0.07	***	(.01)	-0.16	***	(.01)	0.06	***	(.01)	0.05	***	(.01)
FRL	-0.14	***	(.01)	-0.19	***	(.01)	-0.11	***	(.01)	-0.16	***	(.01)
Schl Math/ELA Prof	0.14	***	(.01)	0.24	***	(.01)	0.16	***	(.01)	0.23	***	(.01)
Year1516	-0.01		(.01)	-0.01		(.01)	0.00		(.01)	-0.01		(.01)
Year1617	-0.01		(.01)	-0.01		(.01)	0.00		(.01)	0.00		(.01)
PACE	0.14	***	(.03)	0.04		(.04)	0.08	*	(.03)	0.10	*	(.04)
PACE*Year1516	0.05	**	(.02)	-0.01		(.02)	-0.03	~	(.02)	-0.02		(.02)
PACE*Year1617	-0.04	*	(.02)	0.01		(.02)	-0.03	~	(.02)	0.03		(.02)
PACE*Prior Achvmt	0.01	*	(.01)	-0.01		(.01)	-0.06	***	(.01)	-0.03	***	(.01)
PACE*IEP	-0.01		(.02)	-0.03		(.02)	-0.03		(.02)	0.02		(.03)
PACE*Male	-0.08	***	(.01)	-0.03	*	(.01)	-0.03	*	(.01)	-0.03	*	(.02)
PACE*FRL	0.01		(.01)	-0.01		(.02)	-0.01		(.02)	-0.04	~	(.02)
Variance components												
Within-Schools	0.60	***	(.0)	0.81	***	(.01)	0.67	***	(.01)	0.89	***	(.01)
Between-Schools	0.02	***	(.0)	0.02	***	(.0)	0.01	***	(.0)	0.02	***	(.0)
wtd. N_Students	68,046			68,194			61,277			61,409		
N_Schools	145			145			92			92		

Note. MIXED command in SPSS with ML estimation; ATE inverse propensity score weights applied as a regression weight. Prior Achvmt=prior achievement; IEP=disability status; FRL=free-and-reduced price lunch status; Schl Math/ELA Prof=percent of students proficient or above in math or ELA in the school (depending on the content area of outcome measure); Year1516=2015-16 school year; Year1617=2016-17 school year; Estimate=parameter coefficient; SE=standard error.

****p*<.001, ***p*<.01, **p*<.05, ~*p*<.10.

RQ#2: Subgroup Analysis

Model I was also used to examine differential effects of treatment based on student-level characteristics such as prior achievement, disability status, gender, and socioeconomic status. Table 6 contains parameter estimates for the differential effects using interaction terms—columns 1 and 5 present results for grade 8 and 11 math, respectively, and columns 3 and 7 present results for Grades 8 and 11 ELA. Students' prior achievement, based on their baseline test scores, tends to show negative differential effects of treatment. That is, PACE students who were relatively lower achieving prior to treatment showed relatively greater differential effects than did those who started relatively higher achieving prior to their district's PACE involvement, although the observed effects are very small (*d*=-0.01 to -0.06), setting all other covariates to sample averages. In general, male students appear to have performed relatively lower than their non-PACE comparison peers, when all other covariates are set to sample averages; however, the observed effects are still very small (*d*=-0.03 to -0.08).

The conditional observed effects for PACE special education students and free-and-reduced price lunch students are very noisy with relatively large standard errors. Given that limitation, PACE special education and free-and-reduced price lunch students appear to have done less well than their non-PACE comparison peers, although the results are very small in magnitude (*d*=-0.01 to -0.04). It should be noted that the share of students falling into these categories was fairly small and findings should be interpreted with caution.

To summarize, these subgroup analyses largely indicate that the effects of PACE treatment do vary by prior achievement and gender with higher prior achieving and male students showing negative differential effects. There is inconclusive evidence regarding differential effects of PACE treatment by disability and free-and-reduced price lunch status. All differential effects are small in magnitude.

RQ#3: Dosage Effects

The effect estimates resulting from Model II (Equation 2) were more nuanced because the analysis models variation in treatment dosage by year and therefore splices the treatment group into multiple analytic categories within year (treat1, treat2, and treat3). This means there is a differential number of students in each group (especially the treat3 group) due to the way the PACE schools/districts entered the pilot. The differential number of students in each dosage group should not bias results as residuals are centered at 0 and the variances are constant across groups. Smaller group sizes do impact analytic power, however, and that is especially true of the dosage effect analyses because the data is spliced into smaller groups.

Table 7 provides parameter estimates and variance components for Model II. Figure 2 illustrates the predicted mean differences in SD units for Grade 8 and 11 math and ELA by treatment dosage and year. Non-PACE students are coded in red. Overall, the findings are relatively consistent with the aforementioned positive effect estimates and effect sizes from Model I except there are some dosage levels where PACE students tend to perform lower than their non-PACE comparison peers. For example, both Grade 8 ELA and math have some negative effect estimates in the 2014-15 school year (treat1), 2015-16 school year (treat2), and 2016-17 school year (treat3). These three treatment groups represent the same four PACE districts that were the first to implement the PACE pilot, but include different cohorts of Grade 8 students each year within those districts. In other words, the four school districts that began implementing PACE at the beginning of the pilot tend to exhibit small negative effects (as measured by the state test) after three years of implementation.

Table 7

Model II Parameter Estimates and Variance Components from Multilevel Models Showing the Conditional Effects of PACE on Grade 8 and 11 Math and ELA Achievement Using Inverse Propensity Score Weights

	Grade 8Math			Grade 8 ELA			Grade 11 Math			Grade 11 ELA		
	Estimate		SE	Estimate		SE	Estimate		SE	Estimate		SE
	(1)		(2)	(3)		(4)	(5)		(6)	(7)		(8)
Intercept	0.16	***	(.02)	0.23	***	(.02)	0.02		(.01)	0.05	**	(.02)
Prior Achvmt	0.73	***	(.0)	0.56	***	(.0)	0.70	***	(.0)	0.60	***	(.0)
IEP	-0.20	***	(.01)	-0.39	***	(.01)	-0.12	***	(.01)	-0.30	***	(.01)
Male	-0.11	***	(.01)	-0.18	***	(.01)	0.04	***	(.01)	0.04	***	(.01)
FRL	-0.13	***	(.01)	-0.20	***	(.01)	-0.12	***	(.01)	-0.17	***	(.01)
Schl Math/ELA Prof	0.16	***	(.01)	0.24	***	(.01)	0.16	***	(.01)	0.23	***	(.01)
Year1516	-0.01		(.01)	-0.01		(.01)	0.00		(.01)	-0.01		(.01)
Year1617	-0.01		(.01)	-0.01		(.01)	0.00		(.01)	0.00		(.01)
Treat1	-0.01		(.09)	-0.06		(.09)	0.03		(.04)	0.06		(.07)
Treat2	-0.05		(.09)	-0.09		(.09)	-0.03		(.04)	0.03		(.07)
Treat3	-0.02		(.09)	-0.03		(.09)	0.02		(.04)	0.10		(.07)
Treat1*Year1516	0.21	*	(.10)	0.11		(.10)	0.05		(.06)	0.00		(.09)
Treat1*Year1617	0.14		(.11)	0.08		(.11)	0.01		(.08)	0.09		(.11)
Treat2*Year1617	0.07		(.09)	0.11		(.10)	0.04		(.06)	0.04		(.09)
Variance components												
Within-Schools	0.60	***	(.0)	0.81	***	(.01)	0.67	***	(.01)	0.89	***	(.01)
Between-Schools	0.02	***	(.0)	0.02	***	(.0)	0.01	***	(.0)	0.02	***	(.0)
wtd. N_Students	68,046			68,194			61,277			61,409		
N_Schools	145			145			92			92		

Note. MIXED command in SPSS with ML estimation; ATE inverse propensity score weights applied as a regression weight. Prior Achvmt=prior achievement; IEP=disability status; FRL=free-and-reduced price lunch status; Schl Math/ELA Prof=percent of students proficient or above in math or ELA in the school (depending on the content area of outcome measure); Year1516=2015-16 school year; Year1617=2016-17 school year; Treat1=one year of PACE treatment; Treat2=two years of PACE treatment; Treat3=three years of PACE treatment; Estimate=parameter coefficient; SE=standard error. Only three interactions between the number of treatment years and Year ID was included in the model specifications because of the rolling cohort nature of the implementation.

***p<.001, **p<.01, *p<.05, ~p<.10

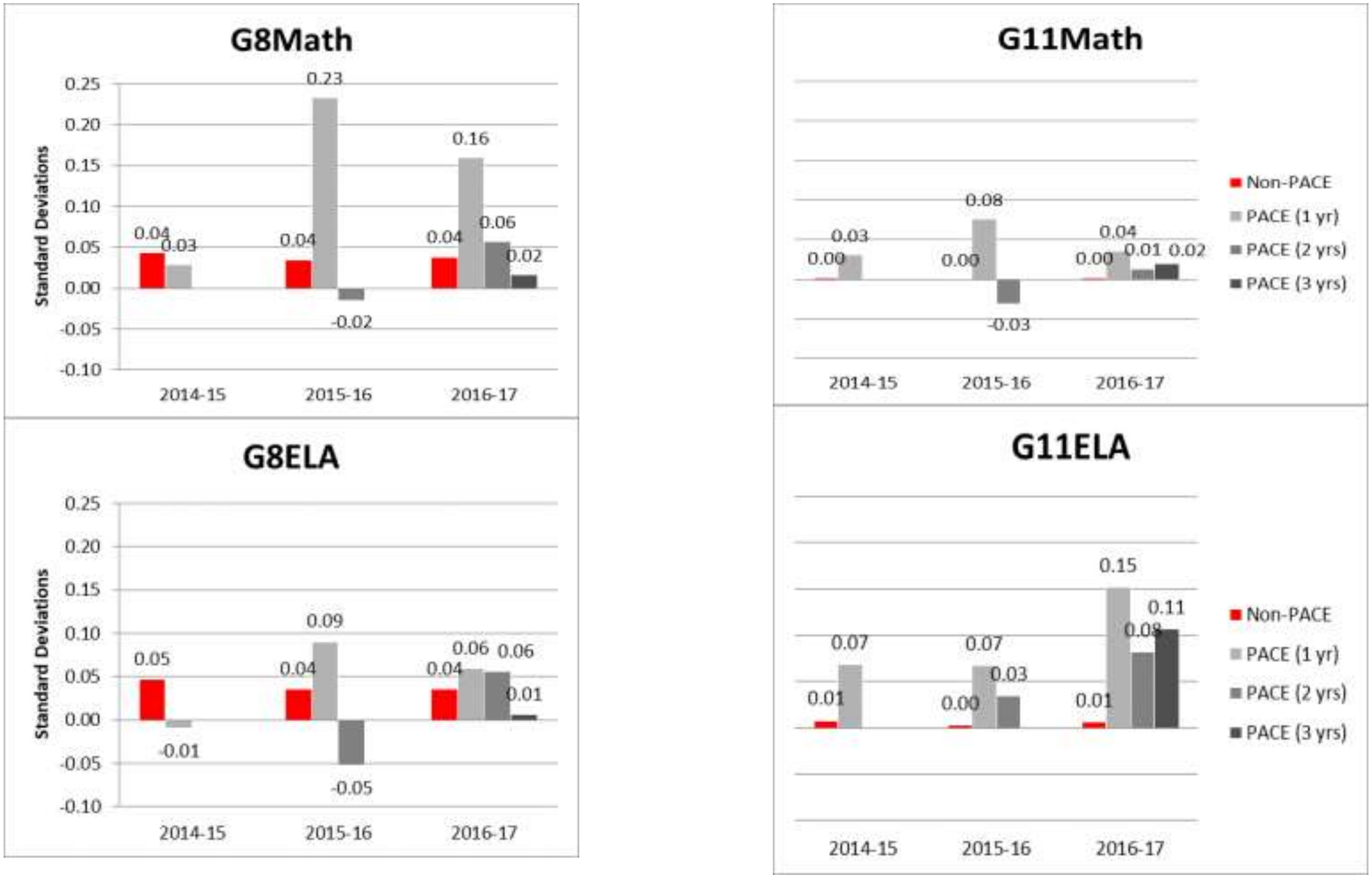

Figure 2. Predicted Conditional Mean Grade 8 and 11 ELA and Math Achievement by School Year and Number of Treatment Years for the Average Student Using the IPSW sample and Model II

Note. Figures illustrate parameter estimates from Model II. Non-significant effects are included. Figure represents the average sample 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 proficient or above in math or ELA, depending on the outcome variable), three treatment variables based on the amount of dosage (one, two, or three years of implementation at the school level), Year ID, and interactions between treatment dosage and Year ID.

There was little to no evidence to suggest that effects accumulate over time the longer a student is exposed to treatment. There was some evidence to suggest either a bump in performance during the first year of treatment or a reduction in performance after the first year of implementation, depending upon how one looks at the results. This pattern of effects may result from the absence of annual test preparation activities in treatment (PACE) districts as students are only exposed to a state achievement test once per grade span in the treatment group. As such, the longer each student has been exposed to treatment is inversely proportional to the number of years without state achievement test preparation activities. This is by no means the only explanation for these findings, but one possible interpretation.

Robustness Checks

The sensitivity of effect estimates to weighting was examined for robustness for all analyses by comparing grade and subject area effects estimated from the IPSW samples with the effects estimated from the unweighted samples for both Model I and Model II. Evidence of selection bias was not found because the average effect estimates between the weighted and unweighted analyses were very similar in magnitude and in the same direction. Also, to check the robustness of the effect estimates, outcome models were fit for each district cohort (districts that started implementing the PACE pilot in the same school year) instead of grouping all district cohorts together. The yearly results were similar to the ones with all cohorts together.

Discussion

It is important to note that this descriptive study's research design rests on large assumptions about baseline equivalence between the treatment (PACE) and comparison (non-PACE) group, suffers from limited power due to a small treatment sample size, and is limited by the available measures. Policymakers and practitioners should be cautious of making any causal interpretations from these early results and instead interpret this study's findings as one source of evidence that the PACE pilot has met the criterion of "no harm" on the state achievement test. This criterion is a key assumption undergirding any attempt to promote innovative assessment and accountability systems and therefore findings from this study may promote the "promise" of alternative systems.

In examining the effect of the PACE pilot on Grade 8 and 11 student achievement outcomes, as measured by the statewide annual achievement test in math and ELA, findings from this descriptive study suggest that there were small positive effects of the PACE pilot in all examined grades and subject areas. These observed conditional effects range in magnitude from about 3% to 14% of a standard deviation. There does not appear to be a consistent pattern of relatively larger effects in math or ELA because effects vary by grade, nor do effects appear to accumulate over time the longer a student is exposed to treatment.

This study's overall findings mirror earlier research on classroom performance assessments. For example, Shepard and colleagues (1995) found similar math gains after one year of a classroom performance assessment interaction (d=0.13) to those shown in Grade 8 and 11 math in this study. 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 five-year time period (Stone & Lane, 2003).

Since it is not uncommon to see little impact on student achievement from major reform efforts during the first three to five years of implementation (Fullan, 2001), the descriptive evidence from this study for a small positive effect for PACE students during the first three 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), the lack of evidence for an implementation dip in this study reflects positively on the PACE pilot.

The findings related to differential effects by subgroup reveal a less consistent picture. Differential effects by subgroup were not examined in the prior literature on the efficacy of performance assessment programs. This study found that lower achieving students tended to exhibit small positive differential effects of PACE treatment. However, there were inconclusive findings related to differential effects for FRL students. There were very small negative differential effects for FRL students in all but one grade and subject area, but all were non-significant. Future research in other state contexts with a larger percentage of economically diverse students is needed to probe and explore these results further.

The final significant interaction effect between a student-level characteristic and treatment status was for male students. Grade 8 and 11 male students tended to exhibit negative differential effects of treatment in both math and ELA. It is unclear why there is a negative interaction between gender and treatment status, especially in math, because previous studies of gender achievement gaps tend to show that males outperform females on math tests, but females outperform males on ELA tests (Fryer & Levitt, 2009; Robinson & Lubienski, 2011). 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). On the 2017 SAT, males tended to score a little lower in ELA (532 versus 534 scale score), but higher in math (538 versus 516 scale score; College Board, 2017). Given that other research suggests that test item format (e.g., multiple-choice tests versus constructed-response questions) explain about 25% of the variation in gender achievement gaps among states on standardized accountability tests (Reardon, Kalogrides, Fahle, Podolsky, & Zarate, 2018), it is unknown if somehow performance-based assessment differentially effects male students. Further research on reform systems that are designed using performance-based assessments could explore these relationships with different populations, subject areas, and grade levels.

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 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. Future research could be conducted in other states and settings with a more ethnically, racially, and linguistically diverse student population to examine effects in those contexts.

Additionally, one challenge in terms of extrapolating from these descriptive findings any conclusions about large-scale performance assessment program reforms is that it is impossible to disentangle the effects of each reform, or other reforms taking place simultaneously within states, districts, and/or schools. For example, NH districts are moving to a more competency-based approach to education as a result of a statewide initiative (NHDOE, 2014a). It is impossible to isolate the effects of the PACE pilot on student achievement outcomes from these other reforms because effects are confounded.

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. Previous research suggests 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), but none of this is accounted for in this study.

There is some implementation commonality across participating districts, however. Specifically, PACE districts co-design the common performance tasks administered across all participating PACE districts, which are used to calibrate scoring within and across districts (NHDOE, 2016b). Furthermore, all PACE districts receive the same capacity building supports, training, and resources from the NHDOE and its technical partners. Future research could conduct research on the levels of implementation fidelity 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.

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 propensity score methods were used 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 after three years of implementation.

Implications for Research, Policy and Practice

The flexibility offered to multiple states under the Innovative Assessment Demonstration Authority to experiment with the use of performance-based, competency-based, interim and/or other types of assessment to determine student proficiency each year is unparalleled since the adoption of NCLB in 2001. This flexibility stems from an interest in the transformation of assessment purpose and use, especially for assessments used in a school accountability context (Baker & Gordon, 2014). Specifically, large-scale performance assessment programs have been forwarded as one solution to close achievement gaps, prepare students for college or career, and facilitate teaching and learning of more cognitively complex competencies (Darling-Hammond et al., 2014). And yet, there is not a lot of empirical evidence on the efficacy of these reforms to improving student achievement outcomes within an accountability context. This study begins to fill the gap in the research base on the efficacy of large-scale performance assessment programs to improving student achievement outcomes in a school accountability context.

However, there is still a lot left to understand about the effects of a so-called "innovative assessment system" such as NH's PACE pilot on student achievement outcomes, as well as many other outcomes. For example, this study focused on Grade 8 and 11 math and ELA student achievement. Future research could explore other outcomes of interest 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). Additionally, this study was not able to establish baseline equivalence, which is important in the future as attention is drawn to the NH model and other states explore innovative assessment systems. Future research could investigate other approaches for establishing such equivalence given the self-selection bias typically inherent in scale-up innovations.

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 the content standards based on the criterion of "no harm" on the state achievement test. Even though effect sizes did not reach commonly acceptable thresholds of practical significance, because differences in performance between the PACE group and the non-PACE comparison group were not statistically significant and practically significant in the direction that favors the traditional assessment mechanisms, there is strong evidence 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 system might look like under the Innovative Assessment Demonstration Authority option, particularly one that utilizes performance-based assessments (Rothman & Marion, 2016). The PACE assessment system design may help address national concerns about over-testing and/or the negative effects of high-stakes testing and accountability on teaching and learning. 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 under the Demonstration Authority, especially as only a few states have indicated their interest in experimenting with their statewide assessment system (Klein, 2018).

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 (Elmore, 2004) rather than external rewards and sanctions to accomplish organizational change and growth. PACE districts are provided capacity building supports and resources from the state to implement the performance-based assessment system and tasked with the responsibility of holding themselves accountability for student growth. For this reason, the PACE system promotes a very different accountability model than *NCLB* where there were specific sanctions faced by schools that did not meet adequate yearly progress and continues to a lesser extent under *ESSA*.

In terms of implications for practice, this descriptive study provides early empirical evidence that learning gains exhibited by students resulting from this large-scale performance assessment reform 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. This also suggests that content coverage in PACE districts is not sacrificed for the sake of content depth as PACE students tended to perform slightly better than predicted on the state achievement test. It will be important to monitor student achievement trends over time as these are early results and the PACE pilot continues to scale to all the districts in the state during the 7-year period of the Demonstration Authority—assuming, of course, that NH's application is approved.

Conclusion

Many schools, districts, and states across the United States pursue assessment and accountability reforms and implement policy changes because of the belief that doing so will improve 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 descriptive 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 provide evidence that PACE students are provided an equitable opportunity to learn the content standards over the first three years of implementation, which fulfills the key criterion of "no harm" on the state achievement test. This research may provide the empirical evidence other states need to move forward with plans to develop innovative assessment systems under the Innovative Assessment Demonstration Authority. Results of this study may also promote new paradigms for assessment and accountability that support deeper learning goals and systemic educational change for all students.

Acknowledgements

The author would like to thank the NHDOE for the use of data, as well as several individuals who provided substantive feedback on earlier drafts, including: Suzanne Graham, Todd DeMitchell, Charlie DePascale, and Emilie Reagan. All errors or omissions are the author's own.

References

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*(5), 258–267. https://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. https://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).

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.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal*, *42*(2), 231–268. https://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. https://doi.org/10.1016/S0742-051X(02)00054-9

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.

College Board. (2017). SAT Suite of Assessments Annual Report. Princeton, NJ: College Board. Retrieved from https://reports.collegeboard.org/pdf/2017-total-group-sat-suite-assessments-annual-report.pdf

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

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. https://doi.org/10.1111/j.1467-9620.2004.00375.x

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. https://doi.org/10.1007/s10464-008-9165-0

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*, *36*(3), 24–34. https://doi.org/http://dx.doi.org/10.1111/emip.12152

Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-based sssessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis*, *20*(2), 95–113. https://doi.org/10.3102/01623737020002095

Fryer, R. G., J., & Levitt, S. D. (2009). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, *2*(2), 210–240. https://doi.org/https://doi.org/10.3386/w15430

Fullan, M. (2001). *Leading in a culture of change*. San Franscisco, CA: Jossey-Bass.

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. https://doi.org/10.1080/00220671.2010.486082

Guo, S., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications* (2nd editio). Thousand Oaks, CA: Sage.

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. https://doi.org/10.3386/w8502

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.

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

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.

Klein, A. (2016). How will ESSA's innovative assessment pilot work? *Education Week Blog*. Retrieved from http://blogs.edweek.org/edweek/campaign-k-12/2016/06/how_will_essas_innovative_asse.html

Klein, A. (2018). States slow to adopt ESSA's testing flexibility. *Education Week Blog*. Retrieved from https://www.edweek.org/ew/articles/2018/01/17/states-slow-to-adopt-essas-testing-flexibility.html

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., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice*, *13*(3), 5–16. https://doi.org/10.1111/j.1745-3992.1994.tb00443.x

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).

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. https://doi.org/10.1207/S15326977EA0804_1

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. https://doi.org/10.1017/CBO9780511815355

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15–21. https://doi.org/10.3102/0013189X020008015

Marion, S., & Leather, P. (2015). Assessment and accountability to support meaningful learning. *Education Policy Analysis Archives*, *23*(9). Retrieved from http://doi.org/10.14507/epaa.v23.1984

McMurrer, J. (2007). Choices, changes, and challenges: Curriculum and Instruction in the NCLB era. Washington, DC: Center on Education Policy.

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal research in educational and social science research* (1st ed.). New York, NY: Oxford University Press.

National Center for Education Statistics. (2015). 2015 mathematics & reading assessments: National scores by student group. Retrieved 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.

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. (2015a). NH PACE progress report (April 30, 2015). Concord, NH: Author.

New Hampshire Department of Education. (2015b). Performance Assessment of Competency Education (PACE). Retrieved 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 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.

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. https://doi.org/10.1080/13803610600696957

Pink, D. H. (2009). *Drive: The surprising truth about what motivates us.* New York, NY: Riverhead Books.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zarate, R. (2018). The Relationship Between Test Item Format and Gender Achievement Gaps on Math and ELA Tests in Fourth and Eighth Grades. *Educational Researcher*, *47*(5), 284–294. https://doi.org/10.3102/0013189X18762105

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. https://doi.org/10.1007/978-94-011-2968-8_3

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal*, *40*(2), 158–189. https://doi.org/10.3102/0002831210372249

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55. https://doi.org/10.1093/biomet/70.1.41

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. https://doi.org/10.1177/0031721716647016

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. https://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.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14. https://doi.org/10.3102/0013189X029007004

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. (1987). Mastery learning reconsidered. *Review of Educational Research*, *57*(2), 175–213. https://doi.org/10.3102/00346543057002175

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).

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. 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., 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). https://doi.org/10.1017/CBO9781107415324.004

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. https://doi.org/10.7249/MG784

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).

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. https://doi.org/10.1207/S15324818AME1601_1

Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership*, (May), 26–33.

About the Author

Carla M. Evans
National Center for the Improvement of Educational Assessment
cevans@nciea.org
https://orcid.org/0000-0001-6250-072X

Carla M. Evans, Ph.D., is a Postdoctoral Fellow at the National Center for the Improvement of Educational Assessment (Center for Assessment). Her research focuses on the impacts and implementation of assessment and accountability policies on teaching and learning. She is interested in policy research related to innovative assessment and accountability systems, competency-based education, performance-based assessments, and teacher/teacher preparation program effectiveness initiatives.

education policy analysis archives

Volume 27 Number 10 February 4, 2019 ISSN 1068-2341

Please send errata notes to Audrey Amrein-Beardsley at audrey.beardsley@asu.edu

Join EPAA's Facebook community at https://www.facebook.com/EPAAAAPE and **Twitter feed** @epaa_aape.