Higher Standards and Lower Achievement?

An Evaluation of Pennsylvania's Keystone Exams

ADAM McGLYNN
East Stroudsburg University of Pennsylvania

As the requirement that students pass three Keystone Exams to graduate from high school was set to take effect, a new round of criticisms arose over the adoption of this assessment system. While the Keystone Exams better align with the new Pennsylvania Core Standards as compared to their predecessor, the Pennsylvania System of Student Achievement, low passage rates especially among low-income and minority students have been cited as a reason to delay implementation of the graduation requirement. This work uses OLS regression analysis to explain which factors are most predictive of school-level performance on the Algebra I, Biology, and Literature Keystone Exams. It finds that race, socioeconomic status, and a school's English Language Learner and special education populations drive performance on the exams. The work concludes by discussing possible policy interventions for the Keystone Exam program going forward.

The impetus for much of the educational reforms that have been enacted over the last 30 years was the need for tougher standards. Quite simply, it was believed that reports such as *A Nation at Risk* (1983) and the results of international assessments such as the Trends in International Math and Science Study (TIMMS) and the Programme for International Student Assessment (PISA) were due to U.S. students not being pushed to achieve

COMMONWEALTH, Volume 18, Issue 1 (June 2016). ISSN 2469-7672 (online). http://dx.doi.org/10.15367/cjppp.v18i1.83.

more. Policymakers believed that if more were required of students, then they would rise to the occasion, and student achievement would increase. Of course, standards are relatively useless unless there is a mechanism in place to assess their achievement. This work focuses on the assessment of the achievement of higher standards in Pennsylvania by examining the recently implemented Keystone Exams, which replaced the Pennsylvania System of Student Achievement (PSSA) in high schools throughout the Commonwealth. Specifically, the question this work intends to address is which factors most influence proficiency rates on the Keystone Exams in Pennsylvania high schools.

There are multiple reasons for addressing this question, but the most important and timely reason is that in February of 2016, Governor Wolf signed legislation delaying the implementation of the requirement that students pass three Keystone Exams in Literature, Algebra I, and Biology in order to graduate from high school. This requirement was planned to take effect during the 2016–2017 school year with the new legislation delaying that requirement by two years until 2018–2019. The reason for the delay was to respond to complaints from school districts over the "high-stakes nature" of the exams and how to help students who have failed the exams complete alternative assessments (Murphy 2016). Therefore, this work will examine demographic, staffing, and programmatic variables in public high schools to identify which factors drive proficiency rates, which in turn could spur a discussion of targeted interventions that would aid high schools in the goal of increasing their proficiency rates on the Keystone Exams.

Background

The PSSA tests mirrored many other high stakes tests in the nation at the high school level where they assessed, for the most part, the minimum level of skill and knowledge believed to be required of a high school graduate. In many cases, the content of the high school PSSA exam was reading and mathematics material that students had learned in eighth and ninth grades, despite the exams being conducted while students were in the 11th grade (Murphy 2015). This means that whether intentional or not, the PSSAs had elements similar to other state assessments and were not necessarily assessing material that students planning to attend college would be exposed to later in high school (Bishop et al. 2000).

The Keystone Exam program thus emanates from a desire to assess student achievement of higher standards adopted in the form of the Pennsylvania Core Standards. The Pennsylvania Core Standards, with which the Keystone Exams are aligned, were derived from the national Common Core State Stan-

dards (CCSS) with the goal of creating "standards aligned with expectations for success in college and the workplace" (Pennsylvania Department of Education 2014, ix). In 2008, the Commonwealth developed the Keystone Exams to create an assessment program "To provide a system that is aligned, focused, standards-based, accurate, universally applicable, and publicly accessible" (Pennsylvania Department of Education 2014, ix).

The Keystone Exams thus serve as "final course exams" (Pennsylvania Department of Education 2014, ix), similar to what one would find in a college-level course. The program was originally charged with creating assessments in 10 subject areas: Algebra I, Algebra II, Biology, Chemistry, Civics and Government, English Composition, Geometry, Literature, U.S. History, and World History. Each exam would account for at least one-third of a student's final course grades, and students who did not score at the level of proficiency or advanced after two attempts would be able to complete a project-based assessment to meet the requirements. However, as of this writing, Pennsylvania has only implemented three exams: Algebra I, Biology, and Literature. These tests are being used to assess high schools as a part of the Pennsylvania School Performance Profile system, which was created under the waiver from the requirements of No Child Left Behind granted to the Commonwealth by the U.S. Department of Education.

Considering the emphasis that has been placed on college attendance in the last 20 years, requiring students to pass final exams that are aligned with a recently completed curriculum makes sense in preparing students for the rigors of undergraduate education. In theory, an end-of-course exam (EOC) would increase standards as it motivates all students to prepare for the test, whereas a minimum-competency test (MCT) will be unlikely to pose a challenge to better performing students (Bishop et al. 2000). Most states adopted MCTs (if they were not already in place) after the adoption of No Child Left Behind in 2002. As NCLB evaluated schools based on a standard of proficiency, minimum-competency tests created an incentive to focus resources on the lowest performing students in a school and a district (Bishop et al. 2000). Achieving high levels of proficiency could ensure that a school and/or a district avoided the punitive measures associated with the failure to achieve adequate yearly progress as required under NCLB, but for the reasons stated above, it does not guarantee a movement toward higher standards and college readiness. Despite these perceived benefits of EOCs compared to MCTs, the United States has little experience with subject-area testing in the assessment of K–12 schools. As Isaacs (2014) notes, SAT subject tests, Advanced Placement (AP) courses and exams, as well as the subject tests conducted by the National Assessment of Educational Progress (U.S. Department of Education 2015) are the most

prominent national examples of subject testing such as occurs in the new Keystone Exam program. It should be noted that none of these assessments form the basis of education policy decisions at the national or state level, meaning that education policy nationally and locally is driven by the results of MCTs.

Therefore, Pennsylvania is limited in the examples it has to draw from in the creation and implementation of its end-of-course exam program. One program that could serve as a guide for the use of EOCs is the New York State Regents Exam program, which has existed for over 100 years. However, for much of the program's existence students were not required to take Regents exams. Students planning to go to a college or university would often complete the Regents track, which required passing five EOCs (Regents exams) in Global History and Geography, English, U.S. History, Mathematics, and Science (Isaacs 2014). The math and science exams are subject specific (i.e., Algebra, Geometry, Chemistry, Biology), but students are only required to pass one from each field. Successfully passing the five exams with a score of 65 or above would earn one a Regents Diploma or a Regents Diploma with honors if a student had an average pass rate of 85 or above. Students not planning to go to college could opt for the Regents Competency Tests (RCTs), which were less rigorous, and could still earn a high school diploma in the process (Isaacs 2014). Less than half of New York State students were graduating with Regents diplomas in the mid-1990s, which led then New York State Commissioner of Education Richard Mills to move to mandate the exams for all students in order to achieve a high school diploma (DeBray 2004).

The expansion of the Regents exam program created many challenges, including concerns over low-income and minority student achievement and the validity of the tests in assessing the state's learning standards, which led to proposed delays in the program's implementation similar to what we currently see in Pennsylvania (Isaacs 2014). Further, there was a concern that standards would need to be lowered to ensure students could pass the exams, which of course in turn could weaken the value of the Regents exam program. This was problematic given that the New York State Education Department released a report that found that state colleges and universities believed only Regents exam scores of 85 or above, what the state viewed as graduation with distinction, were indicative of readiness for higher education (Isaacs 2014; New York State Department of Education 2011). Teachers, while seeing benefits to changes in the testing program, also raised concerns regarding the implementation of testing and corresponding curriculum changes without significant input from teachers (Grant 2000). Isaacs concluded her work by emphasizing that the strength of the Regents program is its alignment with the curriculum, which has only recently come into vogue in other states because of Common

Core. However, this brief discussion of the New York State Regents exams shows that a comprehensive EOC system worked best when it was optional and helped raise standards for above average students; once this system was expanded to all students, significant challenges and growing pains emerged as is the case now in Pennsylvania.

In theory EOC exams are a better assessment tool than an MCT, so why then in an era where there has been greater demand for accountability and increased standards would the Keystone Exams, which seemingly accomplished both, be met with resistance? The reasons are numerous but start with an overall movement against increased testing in schools. In a letter to State Senator Andrew Dinniman, who stands in ardent opposition to the exams, superintendents from more than 50 schools in Southeastern Pennsylvania explained that the exams would place a tremendous burden on high school students and could lead to students spending as many as three or four weeks taking tests when the Keystones are added to AP and local course exams. Additionally, they explained that many eighth graders enrolled in Algebra I would be required to take both the eighth-grade PSSA exams and the Algebra I Keystone Exam (Coalition of Pennsylvania Public School Superintendents and Intermediate Unit Executive Directors 2013). Just as significant to the superintendents was the cost of preparing students for the exam and providing remediation in the form of the project-based assessments that students could complete in lieu of passing a Keystone Exam. In a position paper posted to his website, Senator Dinniman argued that the costs associated with preparing students for the exams, especially those taking it more than once, in addition to the cost of the project-based assessments was an estimated $300 million. He argued that the result is essentially an unfunded mandate placed on school districts by the state, which could ultimately lead to higher property taxes (Dinniman 2013).

Concern over the Keystone Exams is not limited to the southeastern part of the state. Superintendents in the Lehigh Valley also expressed concern about the exams specifically over the project-based assessments. Overall, school leaders in the region were concerned that the project-based assessments have not been clearly explained and would take longer than the eight to ten hours suggested by the state. This would require students to miss class time, staff to be reassigned, and possibly the need for additional staff at a time when local residents are opposed to property tax increases and when state funding has been stagnant (Palochko 2015). Of course, it is unlikely that the Keystone Exams would be met with such vitriol if students were performing admirably on the assessments. However, when the exams were first piloted widely in 2011, only 49.9% of students in the Commonwealth scored at or

above the proficient level on the Literature exam, with 38.6% and 35.7% of students achieving proficiency on the Algebra I and Biology exams, respectively.

This has led to a concern from many groups, including the NAACP (*PennLive* Editorial Board 2009) and school superintendents, that "an inevitable outcome of tying Keystone Exams to graduation rates is an increased dropout rate that will affect a disproportionate amount of low-income and at-risk students than their peers in more affluent communities whose parents and schools have more resources to focus on remediation" (Coalition of Pennsylvania Public School Superintendents and Intermediate Unit Executive Directors 2013). The Keystone Exams are now beyond the pilot phase, so the practicality of the assessments must be evaluated as we approach full implementation. One of the factors that will be examined in this work is whether socioeconomic status and racial and ethnic minority populations drive Keystone Exam proficiency rates as feared by these school and community leaders. These concerns are significant as some members of the Pennsylvania General Assembly, while supporting the purpose of the exams, have called for removing the requirement that the exams be passed in order to graduate from high school (Pennsylvania House of Representatives Education Committee 2015).

As mentioned in the introduction, the Keystone Exams graduation requirement has been delayed until the 2018–2019 school year, which serves as a significant impetus for this research. The delay appears to be an example of where politics is conflicting with what policy analysts and government agencies believe is the best path forward. In response to the proposed two-year delay, the State Board of Education Chairman Larry Wittig, who helped develop the Keystone Exam system, was skeptical. In comments to Penn Live reporter Jan Murphy (2015), Wittig cited the use of the tests as an evaluation measure for teachers as a reason to oppose the delay, but also said, "This was a very well-crafted system. I don't think people critical of it fully understand the ramifications of the meltdown. . . . In the bottom of my heart, I don't feel that by eliminating this requirement, it's doing a positive thing for students" (Murphy 2015). However he also noted that, "If they can come up with something better that achieves the same desired result, I'm all in" (Murphy 2015). It is interesting to note that under the newly passed Every Student Succeeds Act (ESSA), states and schools are no longer mandated to assess teachers using test scores. (The U.S. Department of Education had been granting states waivers from this requirement of No Child Left Behind, in any event.) Therefore, this provides an additional opportunity for policymakers to walk back from the Keystone Exam program if they choose to do so, as they will not have to worry about the pressure to satisfy a federal teacher-evaluation mandate. However, some type of testing must remain—

whether it's the Keystones or something else—as the testing mandate for high school has remained a part of the ESSA.

The move to delay the implementation of the graduation requirement and changes to education policy at the federal level provide an opportunity for policymakers to re-evaluate the program to assess not only the appropriateness of the learning standards and the exams designed to assess them, but the feasibility of future student success in the Keystone Exam program as currently constructed. An educational-input regression model was estimated to assess the factors that appear most influential in predicting school-level success on the Keystone Exams with the goal of helping to identify how policymakers can allocate resources going forward to achieve the goals of the Keystone Exam program.

Data, Methods, and Hypotheses

In November of 2015, the Pennsylvania Department of Education released the results of the Keystone Exam administration for the 2014–2015 Academic Year, representing the most recent available data to assess school performance on the exams. These data provide the dependent variables in this study. Overall, six dependent variables are assessed using Ordinary Least Squares (OLS) regression analysis. Each of the six regression models was estimated using the Stata software program with robust standard errors to correct for the possibility of heteroscedasticity impacting the validity of the regression coefficient estimates. The six dependent variables are the percentage of students achieving proficiency (i.e. passing) the Keystone Exams in Algebra I, Biology, and Literature in each public high school in the Commonwealth and the percentage of Historically Underperforming (HU) students achieving proficiency on the same three exams in each high school in the Commonwealth.[1] As explained by the Pennsylvania Department of Education (PDE) (2015), "The Historically Underperforming group consists of students who are: (1) economically disadvantaged, (2) English Language Learners, or (3) have an Individualized Education Plan. If a student is in more than one of those categories, that student is only counted one time." As the focus of federal education programs has been to improve the achievement of these HU students, it is important to look at Keystone Exam results for both groups. The use of these data provide an advantage over previous analyses of the Keystone Exam program given that student performance improves over time regardless of student ability as students learn strategies specific to a given assessment. As the Keystone Exams were first field tested in 2009 and 2010, and implemented widely in 2011, students and teachers should have familiarity with the assess-

ments at this point. As such, student performance in 2014–2015 should be higher than previous years.

The data for the independent variables were obtained from the PDE Fast Facts data set, which provides information on everything from school addresses to demographic data. For this work the following independent variables are included in the models: school type (charter school or traditional public school), the percentage of teachers in a school who are highly qualified, the average number of years of experience of teachers in the school, the percentage of students in a school who are female, economically disadvantaged, Hispanic, African-American, English Language Learners, identified as gifted, and have an Individualized Education Plan (IEP). Finally, the number of students enrolled in the school and the number of AP courses offered by the school are also included as independent variables.[2] The only public schools omitted from the models are those with missing data and the career, vocational, and technical schools in the Commonwealth often referred to as Career and Technical Centers (CTC) or Area Vocational and Technical Schools (AVTS). These schools' student populations are often composed of students from multiple districts, with many only attending part-time and being assessed by their home districts. This meant that many of these schools had very low testing populations. For these two reasons, they were excluded from the analysis due to concerns that their unique nature could confound the results. After accounting for missing data, 634 high schools in the Commonwealth formed the basis for this analysis with just under 10% of those schools being charter schools.

The purpose of this work is to assess achievement on the Keystone Exams and to help policymakers identify areas of intervention that could help make the program successful. As an alternative, should the results show that achievement gaps based on wealth, race, and ethnicity persist under the Keystone Exam system, policymakers will need to ask whether the program actually benefits students and provides them with an opportunity to be successful. With this goal in mind, this work will test the following hypotheses:

H1: Race, Ethnicity, English Proficiency, Special Education Populations and Socioeconomic Status will impact Keystone Exam proficiency rates negatively in all three subject areas consistent with national performance data on standardized tests (U.S. Department of Education 2015).

H2: Individual school/building level factors such as the gender composition of the student population, the percentage of gifted students, teacher experience, and the number of AP courses offered

will not impact school-level performance on the Keystone Exams. One would expect that having experienced teachers, many gifted students, and more AP courses offered would likely be an indicator of higher proficiency rates on the Keystone Exams; however, there is little variation on the teacher experience and gifted student variables, and many schools offer AP exams as an indication of rigor, but that does not necessarily mean students are earning AP credit, which is why none of these three variables are hypothesized to have an impact.

H3: Schools with lower enrollments will have higher proficiency rates on the Keystone Exams given research that has demonstrated that small schools can aid in overcoming the impacts of socioeconomic status in educational achievement. (Bickel et al. 2001)

Results

Table 1 presents the descriptive statistics for the variables that are included in the regression models. These figures demonstrate a wide disparity in student achievement and the demographics of the student population. We see that while in most schools a majority of students are proficient on all three Keystone Exams, we also see that a majority of the historically underperforming student populations in Pennsylvania high schools do not achieve proficiency in Algebra I and Biology. Also of note is that most schools have highly qualified and experienced teachers, which means those factors are unlikely to be demonstrated to be impactful in explaining school-level performance on the Keystone Exams. Finally, unlike the national trend, African-Americans comprise a much greater proportion of the Pennsylvania student population, with the average high school having an African-American student population of almost 15%, which is more than double the average high school's Hispanic student population.

A few key findings are evident from the regression models in Tables 2 and 3. First, all of the models aid in the explanation of school-level performance on the Keystone Exams although the HU models explain less of the variance in proficiency rates as evidenced by their lower r-squared values.[3] Overall, there is mixed support for all three hypotheses.[4] The most striking finding is that in all six models, the percentage of students in special education programs, the percentage of African-American students, and the percentage of English Language Learners (ELL) had a negative impact on proficiency rates. These significant variables exhibited the lowest p-values, demonstrating the high likelihood that the relationship between test scores and a school's special edu-

Table 1. Descriptive Statistics			
Variable	**Statistic**	**Minimum**	**Maximum**
AVG Proficiency Rate—Literature	72.12	7.58	100
AVG Proficiency Rate—Algebra I	63.70	3.88	100
AVG Proficiency Rate—Biology	57.50	0	100
AVG Proficiency Rate—Literature (HU Students)	58.18	7.69	100
AVG Proficiency Rate—Algebra I (HU Students)	48.47	4.05	100
AVG Proficiency Rate—Biology (HU Students)	*42.64*	0	100
Number of Charter Schools & PCT of Sample	63 (9.94%)	—	—
Percent African-American	14.65	0	100
Percent Hispanic	6.75	0	93.59
Percent Economically Disadvantaged	43.81	3.09	100
Percent English Language Learners	1.80	0	46.44
Percent Female	49.00	0	100
Number of AP Courses Offered	8.16	0	34
Percent Gifted Students	4.1	0	51.86
AVG Number of Students Enrolled	892.18	97	9344
Percent Special Education Students	14.96	0	44.67
AVG Number of Years of Teaching Exp.	13.04	1.83	20.42
Percent Highly Qualified Teachers	97.64	0	100
Note: Statistics based on a sample size of 634, except for the HU proficiency rates where $N = 620$.			

cation population is not due to chance. Not only was the level of significance impressive, but also the magnitude of their impact, especially with the ELL and special education student variables. On average across the six models, a percentage point increase in a school's special education population would yield a percentage point decrease in proficiency rates, while percentage point increases in the ELL population would on average decrease proficiency rates by approximately seven-tenths of a point. Comparatively, although consistently statistically significant, a percentage point increase in the African-American student population would decrease proficiency rates by two-tenths of a point in high schools in the Commonwealth. Thus while race matters in explaining Keystone Exam proficiency rates, the steepest declines will be observed with increases in a high school's ELL and Special Education populations.

Also significant across multiple models was the percentage of students who were economically disadvantaged, which was found to reduce proficiency rates by an average of two-tenths of a point in the total student population models. In an unexpected finding, the percentage of economically disad-

Table 2. Keystone Exam Regression Model for Total High School Student Populations

Variable	Literature (*N* = 634)	Algebra I (*N* = 634)	Biology (*N* = 634)
	Coefficient (S.E.)	*Coefficient (S.E.)*	*Coefficient (S.E.)*
Charter School	–2.40 (2.94)	–8.20 (2.80)**	–6.86 (2.48)**
Percent African-American	–0.15 (0.036)***	–0.218 (0.036)***	–0.255 (0.035)***
Percent Economically Disadvantaged	–0.181(0.039)***	–0.205 (0.044)***	–0.342 (0.042)***
Percent English Language Learners	–0.903 (0.129)***	–0.643 (0.152)***	–0.641 (0.150)***
Percent Female	0.339 (0.113)**	0.211 (0.113)	0.152 (0.108)
Percent Hispanic	*–0.003 (0.067)*	–0.054 (0.066)	–0.059 (0.063)
Number of AP Courses Offered	0.364 (0.076)***	0.420 (0.086)***	0.313 (0.083)***
Percent Gifted Students	0.380 (0.107)***	0.409 (0.128)**	0.299 (0.158)
Number of Students Enrolled	–0.0007 (0.0006)	–0.0015 (0.0006)*	–0.001 (0.001)
Percent Special Education Students	–0.990 (0.125)***	–0.997 (0.108)***	–0.695 (0.093)***
Average Number of Years of Teaching Exp.	–0.222 (0.201)	–0.286 (0.219)	–0.321 (0.216)
Percent Highly Qualified Teachers	0.138 (0.074)	0.159 (0.082)	0.087 (0.079)
Constant	67.80 (10.32)***	67.25 (10.45)***	73.86 (9.89)***
R-squared value	0.656	0.669	0.738

*$p < 0.05$ **$p < 0.01$ ***$p < 0.001$

vantaged students had a small positive impact on the Algebra I proficiency rate but only in the Historically Underperforming student model. This is an anomaly that is difficult to explain but could possibly be attributed to the use of school-level data for this subset of students as discussed previously. Overall, the analysis demonstrates as expected in the first hypothesis that race, socioeconomic status, English language proficiency, and special education status negatively impact student performance. Also worth noting is that there was an independent impact for race and socioeconomic status. One could posit that the disproportionate poverty African-Americans live in compared to Caucasians (Macartney, et al. 2013) could mitigate the impact of race, and only show an impact for socioeconomic status, but here it is appar-

Table 3. Keystone Exam Regression Model for Historically Underperforming Student Populations

Variable	Literature (*N* = 620)	Algebra I (*N* = 620)	Biology (*N* = 620)
	Coefficient (S.E.)	*Coefficient (S.E.)*	*Coefficient (S.E.)*
Charter School	3.03 (3.27)	–3.45 (3.38)	–3.10 (3.05)
Percent African-American	–0.126 (0.039)**	–0.194(0.046)***	–0.244 (0.037)***
Percent Economically Disadvantaged	0.069 (0.040)	0.104 (0.051)*	–0.051 (0.050)
Percent English Language Learners	–0.813 (0.139)***	–0.603 (0.140)***	–0.593 (0.137)***
Percent Female	0.414 (0.115)	0.332 (0.119)**	0.168 (0.115)
Percent Hispanic	*–0.026 (0.072)*	–0.079 (0.066)	–0.092 (0.065)
Number of AP Courses Offered	0.332 (0.102)**	0.354 (0.113)**	0.179 (0.109)
Percent Gifted Students	0.610 (0.125)***	0.574 (0.292)***	0.390 (0.248)
Number of Students Enrolled	–0.001 (0.0007)*	–0.002 (0.0001)**	–0.001 (0.001)
Percent Special Education Students	–1.14 (0.124)***	–1.10 (0.127)***	–0.832 (0.108)***
Average Number of Years of Teaching Exp.	–0.281 (0.252)	–0.247 (0.284)	–0.443 (0.271)
Percent Highly Qualified Teachers	0.120 (0.075)	0.142 (0.084)	0.093 (0.074)
Constant	43.05 (10.30)***	34.65 (11.57)**	49.24 (10.09)***
R-squared value	0.400	0.391	0.476

*p < 0.05 **p < 0.01 ***p < 0.001*

ent that African-American students are struggling to achieve proficiency on the Keystone Exams even with the economically disadvantaged variable in the model.

Additionally, the number of AP courses offered in the high school had a significant and positive impact on proficiency levels in all three models for the total student population and for the Literature and Algebra I models among Historically Underperforming students, contradicting the second hypothesis. Each additional AP course offered is estimated to increase proficiency rates by approximately 0.35 points. Further, the percentage of gifted students was found to increase proficiency rates on the Literature and Algebra I Keystone

Exams: A percentage point increase in the gifted population increases proficiency rates by approximately one-half of a percentage point. The findings related to AP course offerings and the gifted student population contradict the second hypothesis and will be examined in the discussion section. The impact of charter schools was limited. Only in the Algebra I and Biology models for all students in a school did being a charter school have a statistically significant impact, and in those instances charter schools had proficiency rates that were eight points lower in Algebra I and seven points lower in Biology. The results also show that percentage point increases in the female population of a high school would on average increase proficiency rates on the Literature Keystone Exam (total student population model) and Algebra I Keystone exam (HU student model) by three-tenths of a point. And finally, smaller schools are found to have a positive impact on proficiency rates for the Algebra I exam in both models, and the Literature exam in the total student population model, supporting the third hypothesis. However, it should be noted that the magnitude of the impact of smaller schools is low.

Just as important as the findings are the non-findings. In all six models, Hispanic ethnicity, teaching experience, and the percentage of teachers who were highly qualified proved statistically insignificant. The insignificance of the teacher variables was expected and supports the second hypothesis, but the findings related to ethnicity are contrary to hypothesis one. Therefore, despite national test results showing ethnicity negatively impacting student achievement, the results here do not support those findings. While it is difficult to offer any definitive explanations, the smaller overall size of the Hispanic population in Commonwealth high schools could be the cause of the observed result.

Discussion

The results presented here do provide insight into how the Commonwealth can save the Keystone Exam program and achieve its original goals of ensuring that students are achieving higher academic standards. However, this work does not allow for a specific diagnosis of the cause of the achievement gap that is confirmed by this work. There are myriad possible reasons for these findings, including test bias, inadequate funding in minority and low-income schools, lower standards and expectations in the classrooms of historically underperforming students, and concerns that out of school experiences can negate what goes on in the classroom. These reasons are all plausible. However, school finances should be the first issue addressed by policymakers. First, the literature on test bias is controversial, with psychometricians appearing

to believe that mean differences in group test scores are unlikely due to bias (Warne et al. 2014). Also, the adoption of the Common Core standards and recent changes to teacher evaluation systems in the Commonwealth (the result of the waiver Pennsylvania received from the requirements of NCLB) show an attempt to address the possibility that teachers expect less from historically underperforming groups, which a 2014 report from the Center for American Progress found to be the case. Finally, there is a large body of research explaining the impact of outside factors on educational achievement, especially poverty (Jensen 2009), but policymakers are limited in their ability to address such problems due to both budgetary and public opinion constraints. These issues deserve much greater discussion than provided here, but for the reasons stated above and because a recent study by the Education Trust (2015) found Pennsylvania to have the third worst funding gap between low poverty and high poverty districts in the country, school finances will be the focus as one possible solution to the Keystone Exam achievement gap.

First and foremost, the Commonwealth must find ways to provide more services and resources to its special education and English Language Learner student populations. For most schools, the special education population will be of greater concern given its size relative to the ELL population. Just as its predecessor No Child Left Behind did, the new Every Student Succeeds Act requires schools to assess 95% of its student population, and because most schools have 15% of their student population or more requiring special education services, at least two-thirds of that group will need to be tested to meet the requirement. In recent years Pennsylvania revised its special education funding formula, which now accounts for not only student headcounts but also employs weighted cost categories to ensure more funding goes to districts with students who cost the most to educate (Commonwealth of Pennsylvania General Assembly 2013). However, schools continue to struggle with the costs of special education services as school funding has remained relatively flat in Pennsylvania, and the revised formula is being applied only to small increments of additional annual funding. As well, it has been found that school districts end up overpaying charter schools to educate special education students who live in their district based on current law. The Pennsylvania Association of School Business Officials (PASBO) estimates this overpayment to be $200 million a year (PASBO 2014). This problem is likely to worsen given the rising number of students that are classified as being on the Autism spectrum, which entitles them to special education services. Thus of fundamental importance to raising Keystone Exam scores will be a comprehensive restructuring of how special education students are funded and serviced in the Commonwealth.

While it is easier to diagnose and formulate plans regarding special education funding and services, policies that could successfully address racial and economic disparities remain elusive. Many reform advocates have campaigned for more charter schools to address racial and socioeconomic achievement gaps, and while several are operating within the Commonwealth in areas with low-income, minority populations, the results here show that charter schools do not perform better than traditional public schools; in the case of the Algebra I and Biology exams, they perform significantly worse. Thus, the Commonwealth's main objective should be to continue interventions in the communities and schools with large low-income and racial minority populations. This starts with greater resources for programs such as early childhood education, which has been one of the few interventions to demonstrate lasting positive results (Karoly et al. 2005). While it is not feasible to attempt to place more gifted students in high schools that lack such students, it is possible to help increase standards by providing the funds to offer more AP courses. At the start of any such initiative in schools with low proficiency rates, few students are likely to earn college credit through these courses. However, exposure to this advanced material will better prepare students for higher education, even if it does not immediately result in college credit. Lastly, it sends the message to students in these high schools that faculty and administrators believe they are capable of achieving more. Districts could also consider having students spend more time on their schooling with after-school and possibly weekend enrichment programs. Parents and communities should also be incorporated more into the education process. These are interventions that have also proven successful (UCLA Center for Mental Health in Schools 2007) as the amount of time students spend outside of school dwarfs the time spent inside the classroom. Out-of-classroom time sometimes undoes what was achieved in the classroom.

The Pennsylvania General Assembly and the governor are currently at odds over the Commonwealth budget; an impasse that has lasted over eight months as of this writing. A principal area of disagreement is revenue generation to be used for increasing funding to public schools. Based on the results of this paper, while one can posit that money is not everything, supplying more resources for Pennsylvania schools is vital at this time. In theory, the Keystone Exams are a worthwhile program as they help to ensure higher standards are achieved while better aligning curriculum with assessment, something that was missing at the high school level under the PSSAs. However, to close with an analogy, if one can high jump five feet, one can't raise the bar six inches and expect increased effort to immediately result in the achievement of the new standard. Just as an athlete would need to train to meet the new standard,

schools and students, especially in areas with large low-income, ELL, special education, and minority student populations, need the time and resources to be able to better prepare for these assessments. As discussed in the introduction, the Keystone Exam program has worthwhile benefits, however, like the recently ended No Child Left Behind legislation, those left with implementing the program are struggling without the resources needed to achieve its goals, and without significant changes in Harrisburg, the program is likely to see achievement gaps persist as evidenced in this work.

NOTES

1. Individual student data would be ideal for the purposes of this paper, but at present no individual-level data have been published by the Pennsylvania Department of Education on its website, which is the reason that school-level data are used.

2. One obvious omission here is the absence of a variable measuring per-pupil expenditures. Data for this variable are only available at the district/Local Education Agency Level. While there is likely to be a correlation between overall spending and spending at the high school, there are several factors that could impact per-pupil spending in each school such as the technology available in and the special education populations of each school. As such, the decision was made to only include variables measured at the school/building level.

3. The HU population models continue to use data at the school level for the independent variables while only examining proficiency for a subset of the school's population. This could account for the less robust results for the HU models. This is not ideal, however given that different variables could be found to drive proficiency rates among the full school and the HU subset, the HU models were included.

4. Please note that while there are correlations between some of the variables, e.g., the percentage of economically disadvantaged students and the percentage of African-American students, an examination of Variance Inflation Factors (VIF) did not show evidence of multicollinearity.

REFERENCES

Bickel, Robert, Craig Howley, Tony Williams, and Catherine H. Glascock. 2001. "High School Size, Achievement Equity, and Cost: Robust Interaction Effects and Tentative Results." *Education Policy Analysis Archives* 9 (40). http://epaa.asu.edu/ojs/article/view/369 (March 13, 2016).

Bishop, John. H., Ferran Mane, Michael Bishop, and Joan Moriarty. 2000. "The Role of End-of-Course Exams and Minimum Competency Exams in Standards-Based Reforms." July 1. (CAHRS Working Paper #00-9). Ithaca, NY: Cornell University, School of Industrial and Labor Relations, Center for Advanced Human Resource Studies. http://digitalcommons.ilr.cornell.edu/cahrswp/88 (September 15, 2015).

Center for American Progress. 2014. "RELEASE: Teachers Have Lower Expectations for Students of Color and Students from Low-Income Backgrounds." October 6. https://www.americanprogress.org/press/release/2014/10/06/98317/release-teachers-have-lower-expectations-for-students-of-color-and-students-from-low-income-backgrounds/ (February 25, 2016).

Coalition of Pennsylvania Superintendents and Intermediate Unit Executive Directors. 2013. "Letter to Senator Andrew Dinniman." November 5. http://www.senatordinniman.com/wp-content/uploads/2013/11/1KeystonesPositionPaper.pdf (November 5, 2015).

Commonwealth of Pennsylvania. 2013. Pennsylvania General Assembly. "Special Education Funding Commission Report." December 11. http://archive.pasbo.org/Special%20Ed%20Commission%20FINAL%20REPORT.pdf (December 8, 2015).

———. 2015. House of Representatives Education Committee. *Presentation on PA State Assessments.* July 29.

DeBray, Elizabeth. 2004. "Richard Mills and the New York State Board of Regents, 1995–2001." *Journal of Cases in Educational Leadership* 7 (2): 27–43.

Dinniman, Andrew E. 2013. "8 Reasons Why We Oppose the Keystone Exams. Minority Chair, Senate Education Committee." http://www.senatordinniman.com/wp-content/uploads/2013/11/EightReasonsAgainstKeystones.pdf (November 5, 2015).

Education Trust. 2015. "Funding Gaps 2015." March 25. https://edtrust.org/resource/funding-gaps-2015/ (February 25, 2016).

Grant, S.G. 2000. "Teachers and Tests: Exploring Teachers' Perceptions of Changes in the New York State Testing Program." *Education Policy Analysis Archives* 8 (14). http://epaa.asu.edu/ojs/article/view/405/528 (November 5, 2015).

Isaacs, Tina. 2014. "150 Years of Statewide Assessment in New York: Are Regents Exams Still Fit for a Purpose?" *Assessment in Education: Principles, Policy and Practice* 21 (3): 344–357.

Jensen, Eric. 2009. *Teaching with Poverty in Mind: What Being Poor Does to Kids' Brains and What Schools Can Do about It.* Alexandria, VA: Association for Supervision and Curriculum Development.

Karoly, Lynn, M., Rebecca Kilburn, and Jill S. Cannon. 2005. "Proven Benefits of Early Childhood Interventions." RAND Corporation. http://www.rand.org/pubs/research_briefs/RB9145.html (December 1, 2015).

Macartney, Suzanne, Alemayehu Bishaw, and Kayla Fontenot. 2013. "Poverty Rates for Selected Detailed Race and Hispanic Groups by State and Place: 2007–2011." *American Community Survey Briefs.* Washington, DC: U.S. Census Bureau. https://www.census.gov/prod/2013pubs/acsbr11-17.pdf (February 29, 2016).

Murphy, Jan. 2015. "Graduation Testing Requirement Delay Wins House Passage." *PennLive,* November 24. www.pennlive.com/politics/index.ssf/2015/11/graduation_testing_requirement_1.html (November 24, 2015).

———. 2016. "Wolf Signs Bill to Suspend Use of Keystone Exams as a Graduation Requirement." *PennLive,* February 3. http://www.pennlive.com/politics/index.ssf/2016/02/wolf_signs_bill_to_suspend_use.html (February 26, 2016).

National Commission on Excellence in Education. 1983. "A Nation at Risk: The Imperative for Education Reform." April. http://www2.ed.gov/pubs/NatAtRisk/index.html (October 29, 2015).

New York State Education Department. Board of Regents. 2011. *Potential Revision of High School Graduation Requirements.* March 11. Albany, NY: NYSED. http://www.regents.nysed.gov/common/regents/files/documents/meetings/2011Meetings/March2011/311p12ccrgradreqpresentation.pdf (November 5, 2015).

Palochko, Jacqueline. 2015. "Keystone Exams One Big Headache for Lehigh Valley School Districts." *The Morning Call,* May 10. http://www.mcall.com/news/local/mc-lehigh-valley-keystone-project-based-assessments-20150509-story.html (November 5, 2015).

PennLive Editorial Board. 2009. "Important Questions Remain about Keystone Exams."

PennLive, August 30. http://www.pennlive.com/editorials/index.ssf/2009/08/important_questions_remain_abo.html (October 29, 2015).

Pennsylvania Association of School Business Officials. 2014. "Misleading Claims by Charter School Advocates Must Not Be Permitted to Stall Needed Special Education Reform." May 13. http://archive.pasbo.org/Special%20Education%20Funding%20Press%20Release.pdf (December 8, 2015).

Pennsylvania Department of Education. 2014. *Keystone Exams: Technical Report.* Harrisburg, PA: Department of Education. http://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/Keystone%20Exams/Technical%20Reports/2014%20Keystone%20Technical%20Report.pdf (October 1, 2015).

———. 2015. *2015 Keystone Exam School Level Data.* Harrisburg, PA: Department of Education. http://paschoolperformance.org/Downloads (November 18, 2015).

UCLA Center for Mental Health in Schools. 2007. "After-School Programs and Addressing Barriers to Learning." http://www.isbe.net/learningsupports/pdfs/after-school-prog.pdf (December 10, 2015).

U.S. Department of Education. 2015. *National Assessment of Educational Progress (NAEP).* Washington, DC: Department of Education. http://nces.ed.gov/nationsreportcard/ (November 29, 2015).

Warne, Russell T., Myeongsun Yoon, and Chris J. Price. 2014. "Exploring the Various Interpretations of 'Test Bias.'" *Cultural Diversity and Ethnic Minority Psychology* 20 (4): 570–582.