Center for Education Evaluation and Policy Analysis
Department of Education Policy Studies
College of Education
The Pennsylvania State University

Problems and Prospects for School Accountability in Pennsylvania

Ed Fuller, PhD
Associate Professor, Department of Education Policy Studies
Executive Director, Center for Education Evaluation and Policy Analysis
The Pennsylvania State University

with
Penn State Department of Education Policy Studies Students:
Chi Nguyen
Andrew Pendola
Joe Levitan

EXECUTIVE SUMMARY

Introduction

School accountability systems have become a fixture of the US education landscape. While there is some evidence that school accountability efforts have had some positive effects on student achievement[i], evidence also suggests that accountability systems have created as many problems as solutions.[ii] Indeed, there is widespread agreement that accountability systems adopted under NCLB, as well as through programs such as Race to the Top and NCLB waivers, had significant flaws. In an effort to remedy these flaws, the new version of ESEA—entitled the Every Student Succeeds Act (ESSA)—provides much greater flexibility to states in developing their school accountability systems. Further, ESSA includes some mandates that push states to address some flaws of the former accountability systems. Most important of these mandates is that states must include some non-academic indicators in their systems.

Thus, state policymakers have the opportunity to create accountability systems that more accurately capture school effectiveness in terms of both student achievement and other important student outcomes not related to test scores. The opportunity to redesign accountability systems also provides state policymakers the opportunity to re-build some of the trust lost between educators and local, state, and federal policymakers over the past decades due to poorly constructed accountability systems and the use of accountability results in educator evaluation systems. Pennsylvania policymakers are already in the midst of developing a new SPP to be called Future Ready PA. This White Paper reviews the basis of state accountability systems, the purposes of such systems, the issues with each metric in the SPP, and the issues with proposed metrics in Future Ready PA. Finally, we propose a new accountability system to spur discussion about what metrics should be included in Future Ready PA and how those metrics should be measured.

Purposes of State Accountability Systems

In designing school accountability systems, policymakers must first identify the goals of such systems. Rather than agree on one over-arching goal, policymakers have generally adopted systems that attempt to serve multiple goals. For example, most accountability systems attempt to: (1) provide clear expectations about the level of student achievement that is expected from each school; (2) set aspirational goals for schools regarding student outcomes; (3) document achievement levels of students in schools; (4) identify the effectiveness of schools in improving educational outcomes of students; and, (5) provide information to the public and policymakers about the performance of each school.

Attempts to concomitantly serve all of these purposes, however, has proved quite problematic because these goals are often in conflict with one another. This conflict arises because of differences in the type of data used and the methods employed to analyze the data necessary to address the different purposes. For example, goals one through three can easily be accomplished with the use of very simple metrics such as the percentages of students scoring proficient or advanced, progressing from one grade to the next, and completing high school within four years. Alternatively, goals four and five require the use of sophisticated statistical techniques that remove the influence of factors outside the control of educators to calculate the improvements in student outcomes over time of each school.

Importantly, calculating outcomes for the first set of measures does not require the removal of the influence of factors outside the control of educators while calculating outcomes for the second set of measures does require the removal of the influence of factors outside the control of educators. Thus, the first set of measures captures both the effectiveness of schools and the factors outside the control of educators that influence outcomes such as student background experiences and school size while the second set of measures theoretically only captures school effectiveness.

Problems with the Current School Performance Profile (SPP)

As I and others have noted, there are a number of serious flaws in the current SPP. One prominent flaw is that the majority of the SPP score is derived from student performance on state tests. Parents, policymakers, and the public have largely agreed that the narrowness of the SPP is a flaw that should be addressed in Future Ready PA.

A more important, albeit less obvious flaw, is that the majority of the SPP score is derived from measures that capture both school effectiveness and the influence of factors outside the control of educators such as student background experiences, school size, school location, expenditures per pupil, community support for education, and numerous other factors. This is a flaw because the inability of the SPP measures to remove the influence of factors outside the control of educators means that the SPP score is not an accurate indication of school effectiveness. In fact, as we will show in this report, many of the SPP measures are capturing the influence of factors outside the control of educators. For example, the majority of the current SPP measures are relatively highly correlated with the percentage of students living in poverty in a school. Thus, some proportion of the SPP score is simply a reflection of the percentage of students living in poverty that are enrolled in a school.

This means that the SPP score is not an accurate or fair indication of school effectiveness and, as we discuss in this report, the inaccuracy and unfairness has serious consequences for students, parents, educators, and communities. In particular, the system has negative consequences for schools and communities with high proportions of students living in poverty, students of color, and low property values. These are the very communities to which the Commonwealth should be providing the greatest assistance but, historically, have been the communities that have received the least assistance from the Commonwealth.

Problems with the Proposed Future Ready PA Index Accountability System

The proposed Future Ready Index (FRI) will likely include a greater number of measures that are not directly connected to student performance on state tests, thus will address the flaw in the SPP of having most measures connected directly to student performance n state tests. However, the proposed FRI appears to continue to rely heavily on measures that capture both school effectiveness and the influence of factors outside the control of educators. As such, ***the FRI will also not be an accurate or fair indication of school effectiveness and the inaccuracy and unfairness will have serious consequences for students, parents, educators, and communities.*** We, in fact, believe that the proposed configuration of the FRI will continue to harm the Commonwealth's schools and communities that are already most disadvantaged.

Suggestions for Improving the Future Ready Index Accountability System

Fortunately, there are several options that the Commonwealth could take to improve the school accountability system. First, the Commonwealth could more heavily weight those measures that are least correlated with student background experiences such as the student growth measures. Second, the Commonwealth could use accepted statistical techniques to remove the influence of factors outside the control of educators to calculate the FRI score. Third, the Commonwealth could allow individual districts to create some measures that would be included in their overall accountability score. Fourth, the Commonwealth could report a score for the overall FRI and also include a second FRI score derived entirely from measures that were statistically adjusted to remove the influence of factors outside the control of educators.

Failure to act to create an accountability system that is accurate and fair will continue to disadvantage the very schools and communities that have long been disadvantaged through an inequitable school finance system. We strongly encourage state policymakers to create a system that is accurate and fair to all schools and communities in the Commonwealth.

I. INTRODUCTION

School accountability systems have, for better or for worse, become a permanent fixture of the education landscape. For example, Pennsylvania has attempted to measure and report the performance of K-12 public schools for decades.[iii] School accountability has been described as an effort that "lifts all boats"[iv] and, in fact, there is some evidence that school accountability efforts have had some positive effects on student achievement.[v] On the other hand, evidence has also revealed that some accountability systems have created as many problems as solutions.[vi]

Despite evidence that state accountability systems have had some small, positive impact on academic achievement in some states, there is widespread agreement that most state accountability systems have significant flaws. Pennsylvania responded to this recognition by originating an effort to review and modify the Commonwealth's school accountability system entitled the School Performance Profile (SPP). Shortly thereafter, the passage of the Every Student Succeeds (ESSA) Act provided states the opportunity to revise their school accountability systems and incorporate changes that address the lessons from the past. More specifically, ESSA sets less stringent guidelines for those designing school accountability systems than in the past, although federal guidelines still substantially limit how states can re-design their systems. For example, ESSA still mandates that states focus heavily on measures of student achievement although ESSA also requires the inclusion of at least one indicator of school quality that is not a direct measure of student achievement.

Thus, under ESSA, state policymakers have the opportunity to create accountability systems that more accurately capture school effectiveness in terms of both student achievement and other important student outcomes not related to test scores. The opportunity to redesign accountability systems also provides state policymakers the opportunity to re-build some of the trust lost between educators and local, state, and federal policymakers over the past decades due to poorly constructed accountability systems and the use of accountability results in educator evaluation systems. This White Paper will summarize what the research says about each discretionary element required by the ESSA, as well as propose some alternatives and recommendations.

A. Review of the Every Student Succeeds (ESSA) Act

ESSA devolves significant decision-making authority back to state education agencies and state legislatures. Perhaps the most critical area in which the federal government returned authority to the states is the design and implementation of state school accountability systems.

While ESSA allows for states to develop their own school accountability systems, ESSA does include some specific provisions about such systems that must be adhered to by states. For example, states must:

- Continue to report results by all required subgroups specified under NCLB, including socio-economic and racial/ethnic subgroups;
- Categorize schools based on state-determined goals and methodology, including the lowest 5% of Title I schools, high schools with four-year graduation rates below 67%, and schools with low-performing sub-groups; and,
- Include five performance areas, one of which must be a non-cognitive measure that meaningfully differentiates schools, provides reliable estimates across years, allows for valid inferences to be made, and can be applied to all grade spans included in the accountability system.

More specifically, ESSA requires states to include the following five areas:

- Academic achievement;
- Another valid and reliable academic indicator (e.g., student growth or achievement gap closure);
- Graduation rate;
- A measure of English language proficiency for English Language Learner students;
- Another non-cognitive indicator of school quality.

Finally, states must have a planned response if a school's test participation rate is lower than 95%.

ESSA legislation includes the following examples of non-cognitive measures: student engagement; educator engagement; student access to and completion of advanced coursework; post-secondary readiness; and, school climate and safety. States, however, are allowed to choose additional measures.

B. Purposes of School Accountability Systems

There are numerous objectives of school accountability systems. In general, these objectives can be separated into two different areas: (1) Expectations and Aspirations; and, (2) Assessment of Effectiveness. We list some of these objectives under each of these two goal areas below.

1. Purpose One: Hold High Expectations and Aspirations

Clearly, one primary goal of school accountability systems is to hold expectations and aspirations and communicate those expectations and aspirations through metrics in the accountability system. Within this broad goal are multiple objectives such as the following:

- Provide clear and measurable expectations for what students know and can do to all schools through the reporting of the percentage of students proficient/advanced;
- Set aspirational goals for student achievement, particularly regarding the completion of high school, completion of advanced coursework, and achievement of college-/career- readiness status;
- Provide evidence about the degree to which schools have provided equitable educational opportunities and outcomes; and,
- Provide parents, the public, and policymakers with an easy to understand metric of the degree to which schools have achieved these goals.

2. Purpose Two: Identify School Effectiveness

An additional goal of school accountability systems is to identify the effectiveness of each school within a state. In so doing, states attempt to address multiple objectives such as:

- Provide accurate information about school effectiveness as a signal to educators about their efforts to improve school outcomes and as an incentive to improve outcomes;
- Provide a measurement of principal effectiveness in improving school outcomes;
- Provide an incentive for teachers to collaborate to improve school outcomes through the inclusion of the SPP score in teacher evaluations;
- Create a holistic and accurate view of school quality and effectiveness for use by parents, the public, and policymakers; and,

- Provide accurate information about school effectiveness to state policymakers that can be used to make decisions about how to improve the entire state system of education and identify schools in need of support and assistance.

This goal is particularly important because ESSA requires states to use accountability systems to identify the lowest performing Title I schools and then provide assistance and sanctions to such schools. Thus, very high-stakes are attached to state accountability systems under ESSA which underscores the need for states to create fair and accurate systems that are defensible.

3. Conflicting Purposes

Unfortunately, when included in school accountability systems, these multiple goals and objectives often conflict with each other. Most importantly, the expectations/aspirations objectives are in substantial conflict with the efforts to identify the effectiveness of schools and educators. Specifically, holding high and aspirational expectations is often best communicated through the setting of performance targets that all schools are expected to meet. For example, a state could set a target that 75% of all students must score proficient or advanced on state mandated tests in order for a school to be considered acceptable. In contrast, identifying school effectiveness is best measured by using sophisticated statistical approaches to calculate the progress a school is making with the students enrolled in the school. So, for example, a state could use a value-added statistical approach to calculate a school's student growth on a state mandated test in mathematics.

Most states—including Pennsylvania—have attempted to address both goals by including a variety of status, growth, and other measures in school accountability systems. It is this attempt to achieve the two goals—particularly through the use of one summative score that designates a school's overall performance—that is at the core of the conflict inherent in school accountability systems and what makes the design of the "perfect" accountability system impossible.

In the next section, we review the Pennsylvania SPP and delve deeper into the origins of this conflict. Specifically, we discuss the different types of data in the SPP and how each data point is calculated. Finally, we review the most current weighting system and examine how this weighting system advantages and disadvantages particular types of schools.

C. Understanding Tables in this Report

In the remainder of this report, we include a number of tables that report on the correlations between indicators included in the SPP and school characteristics, including school size and student characteristics.

Correlations vary from a low of -1.0 to a high of 1.0. The further away from 0.0 the correlation coefficient, the stronger the relationship between the two variables. Thus, a correlation coefficient that approaches -1.0 would indicate a very strong negative correlation between two variables while a correlation coefficient that approaches +1.0 would indicate a very strong positive correlation between two variables. Correlation coefficients near 0.0 would indicate extremely weak correlations between two variables.

We use red and green shading to indicate the statistically significant relationships and the strength of those relationships. Cells shaded in white are not statistically significant or have a statistically significant correlation, but the correlation coefficient is between -0.199 and +0.199. Such correlation coefficients reveal the relationship between two variables is extremely weak which is why we have chosen to not shade such relationships.

The lightest shading of red indicates a statistically significant, but relatively weak relationship. The moderate red shading indicates a statistically significant relationship that is

moderately strong. The darker red shading indicates a statistically significant relationship that is strong.

The light green shading indicates a positive and statistically significant correlation that is relatively weak, but still of import. The moderate green shading indicates a positive and statistically significant correlation that is moderately strong. The darker green shading indicates a positive and statistically significant correlation that is strong.

II. REVIEW OF THE PENNSYLVANIA SCHOOL PERFORMANCE PROFILE

In this section, we begin by reviewing the current configuration of the Pennsylvania School Performance Profile (SPP). Included in this initial overview is a discussion of the broad problems associated with the current configuration of the SPP. Subsequently, we review each type of measure included in the current configuration of the SPP. Within these reviews, we focus on some of the weaknesses of the various measures as related to the ability of the SPP to accurately identify the degree of effectiveness to improve student outcomes of Pennsylvania schools.

A. Current Configuration of the SPP

Table 1 includes the SPP indicator areas and the individual indicators within each area. In addition, the table includes the factor value for each indicator. This factor value indicates the weight for each indicator. The current SPP includes five indicator areas: Indicators of academic achievement, Indicators of closing the achievement gap - All students, Indicators of closing the achievement gap - Historically underperforming students, Indicators of academic growth/PVAAS, and Other academic indicators. The two areas with the greatest total factor values are indicators of academic achievement and indicators of academic growth/PVAAS—both with factor values of 40 points.

Table 1: Current Indicators and Point Values for the Pennsylvania School Performance Profile

Accountability Indicator Areas and Individual Indicators	Factor Value EL/MS	Factor Value HS
I. Indicators of Academic Achievement	**40**	**40**
a. Mathematics/Algebra I - Percent Proficient or Advanced on PSSA/Keystone	7.5	7.5
b. ELA/Literature - Percent Proficient or Advanced on PSSA/Keystone	15.0	15.0
c. Science/Biology - Percent Proficient or Advanced on PSSA/Keystone	7.5	7.5
d. Industry Standards-Based Competency Assessments - % Competent or Advanced		5.0
e. Grade 3 Reading - Percent Proficient or Advanced on PSSA	10	
f. SAT/ACT College Ready Benchmark		5.0
II. Indicators of Closing the Achievement Gap - All Students	**5**	**5**
a. Mathematics/Algebra I - Percent of Required Gap Closure Met	1.25	1.25
b. ELA/Literature - Percent of Required Gap Closure Met	2.5	2.5
c. Science/Biology - Percent of Required Gap Closure Met	1.25	1.25
III. Indicators of Closing the Achievement Gap - Historically Underperforming Students	**5**	**5**
a. Mathematics/Algebra I - Percent of Required Gap Closure Met	1.25	1.25
b. ELA/Literature - Percent of Required Gap Closure Met	2.5	2.5
c. Science/Biology - Percent of Required Gap Closure Met	1.25	1.25
IV. Indicators of Academic Growth/PVAAS	**40**	**40**
a. Mathematics/Algebra I - Meeting Annual Academic Growth Expectations	10.0	10.0
b. ELA/Literature - Meeting Annual Academic Growth Expectations	20.0	20.0
c. Science/Biology - Meeting Annual Academic Growth Expectations	10.0	10.0
V. Other Academic Indicators	**10**	**10**
a. Cohort Graduation Rate		2.5
b. Promotion Rate	5.0	
c. Attendance Rate	5.0	2.5
d. Advanced Placement, International Baccalaureate Diploma, or College Credit		2.5
e. PSAT/Plan Participation		2.5
TOTAL BASE POINTS	**100**	**100**
VI. Extra Credit for Advanced Achievement	**7**	**7**
a. Percent PSSA/Keystone Advanced - Mathematics/Algebra I	1.0	1.0
b. Percent PSSA/Keystone Advanced - ELA/Literature	2.0	2.0
c. Percent PSSA/Keystone Advanced - Science/Biology	1.0	1.0
d. Percent Advanced - Industry Standards-Based Competency Assessments	1.0	1.0
e. Percent 3 or Higher on an Advanced Placement Exam	2.0	2.0
TOTAL POSSIBLE POINTS	**107**	**107**

B. Overarching Issues with Current Configuration of the SPP

Historically, school accountability systems have included both cognitive measures (student achievement) and non-cognitive measures (promotion, attendance, and graduation) of student performance. Most systems have emphasized test-based measures of cognitive performance either through the exclusion of other measures or by applying greater weights to test-based measures than to other measures. Much of this focus was driven by federal policy mandates through No Child Left Behind (NCLB), NCLB Waivers, or Race to the Top (RttT) that left states relatively little flexibility in designing their state accountability systems.

As the Pennsylvania Department of Education (PDE) has acknowledged, there are a number of problems with the existing configuration of the SPP. Indeed, PDE has conducted meetings across the state to gather educator feedback as part of the process to create a replacement for the SPP that will be called the “Future Ready PA Index”.

1. Extreme Focus on Test Score Results

The first overarching issue is that the current SPP focuses almost exclusively on measures that stem directly from student assessment results on the Pennsylvania System of School Assessment (PSSA) and Keystone examinations. In fact, 90% of the base SPP score for elementary schools and middle schools is derived from indictors derived from PSSA and Keystone scores. At the high school level, 80% of the base SPP score is derived from indictors derived from PSSA and Keystone scores. Educators from across the Commonwealth expressed deep concerns about the narrow focus of the SPP and urged PDE to include additional indicators that were not connected to PSSA or Keystone test scores.

2. Validity Issues

Evaluations such as school accountability programs should be based on defensible criteria[vii] that lead to "ethical, fair, useful, feasible, and accurate" conclusions.[viii] This recommendation is also referred to as construct validity—the ability of the evaluation effort to provide accurate information that can lead educators and policymakers to make appropriate conclusions from the information. The *Joint Committee on the Standards for Educational Evaluation* recommends that only evaluations that can provide evidence about construct validity should be used.

In order for the signaling effects of the SPP scores to have their intended effect on educators, educators must perceive the system as accurate, fair, and equitable.[ix] In other words, the SPP scores must have "face validity" [x] from the perspective of educators. If, in fact, educators *do not* perceive the SPP scores to have face validity, then they are likely to ignore, subvert, or "game" the entire evaluation system.[xi]

If an evaluation such as a school accountability system lacks either construct or face validity, then the system will clearly not have the intended effects upon educators. Thus, it is critical that the Commonwealth provide evidence about the construct validity of the SPP effort in order to ensure that (1) educators view the SPP scores as credible and (2) use them in ways that effectively improve schools.

Given that the face validity of the SPP is likely to be largely dependent on the perception of the construct validity of the system, the second overarching issue with the SPP is construct validity. In section C below, we review each type of measure included in the current SPP and, within the review, examine the construct validity of the specific indicators within each type of measure.

C. Review of Construct Validity of the Types of Measures Included in the Current SPP

While there are five primary indicator areas in the current SPP, there are three types of measures: status measures, growth measures, and achievement gap measures. Status measures assess student outcomes or perceptions at a single point in time. In short, such measures provide information about the current status for a specific indicator. Growth measures, alternatively, assess changes in outcomes or perceptions over time. In most instances, these growth measures attempt to assess changes in student achievement over time. Finally, achievement gap measures typically assess the differences in achievement levels between various sub-populations of students. In the current configuration of the SPP, the two groups of students are not historically underperforming students and historically under forming students. According to PDE, "The historically underperforming student group is a non-duplicated count of students with disabilities with an individualized education plan (IEP), students who are English Language Learners (ELL), and Economically Disadvantaged (ED) students."

1. Status Measures

Status measures have typically included simple percentages of students meeting a specified standard for a particular test in a particular year. For example, the SPP includes status measures that report on the percentage of students scoring proficient or advanced. Status measures address the expectations and aspirations purposes of school accountability systems by providing evidence about the percentage of students exhibiting the knowledge and skills determined to be appropriate by state policymakers. Such measures can potentially provide useful information about the knowledge and skills of students under certain assumptions.[1]

Rather than assume that test results provide useful information, the Commonwealth should be responsible for providing evidence about the degree to which the results allow for valid and reliable inferences to be made from the test scores. Thus, the Commonwealth should conduct studies to ensure that the adopted cut points allow for valid conclusions to be made about the student's individual test score as well as valid conclusions about the school. In both cases, the separate studies would need to show that the cut points utilized by the Commonwealth to make judgments about students and schools reflect educationally meaningful distinctions.

Most importantly, the Commonwealth should conduct predictive validity studies that examine the degree to which scoring in the various levels (Below Basic, Basic, Proficient, and Advanced) predict other cognitive and non-cognitive outcomes such as other test scores, advancement to the next grade level, grade point average, graduation, and/or success in college. For example, PDE could conduct a study to determine if the Keystone scores are predictive of success in college coursework. If students scoring just below and just above the proficient cut point on the Keystone Algebra I test were equally likely to enroll in post-secondary institutions of education and were equally likely to earn a grade of "B" or greater in a college mathematics course, then the cut point associated with the identification of proficient students would not have predictive validity with respect to college readiness.

Even if such studies document the ability of the tests to yield valid and reliable information, the reliance on the percentages of students scoring at specific levels remains problematic. Most importantly, the reliance on the percentage of students scoring greater than a particular cut point is that such percentages do not provide information about the full range of school-level achievement. For example, suppose two schools—School A and School B--have an identical 60% of students scoring proficient or advanced as reported in the SPP. Many individuals would conclude that the two schools have equivalent levels of achievement. Yet, the overall average scale score for School A could be substantially greater than the overall average score for School B. Thus, even though School A had a substantially greater overall average scale score than School B, the schools would appear to have equal achievement based only on the percentage of students scoring proficient or advanced. The use of four score groupings in Pennsylvania's reporting of results somewhat mitigates this problem, although the SPP includes only two groups. Access to overall average scale scores and standard deviations for each school would provide additional useful information about the actual achievement in each school and grade level and should be available in the SPP or from the DPDE website.

Another important problem with a status measure such as the percentage of students scoring proficient or advanced is that the measure only incentivizes a school to ensure each child achieves

[1] These assumptions include that the test has been constructed in an appropriate fashion to meet established psychometric properties. More importantly, this statement assumes that teachers have not focused on "teaching to the test" in ways that are considered inappropriate by testing experts. For example, efforts by teachers to review specific questions thought to be similar to the questions likely to be on the test is inappropriate because such efforts corrupt the test scores such that they no longer allow valid and reliable inferences to be made about what students know and can do.

at a particular level or above. Indeed, status measures provide absolutely no incentive for schools to ensure students already scoring proficient or advanced make academic progress. This has led schools to place a greater focus on "bubble" students—those scoring just below the specified level—than on other students.[xii] Importantly such an incentive works against the goal of improving the achievement of all students. Moreover, such an incentive is associated with "teaching to the test" strategies in an effort to push students "over the bar" rather than ensuring that students learn the full curriculum.[xiii] Research has consistently shown that such "teaching to the test" strategies has two negative outcomes: students don't actually learn the knowledge and skills embedded in the curriculum and the student scores are not accurate indicators of what students know and can do.[xiv]

Finally, one of the biggest drawbacks of status measures is that research has consistently found them to be highly correlated with student demographics and other school characteristics. In particular, almost all status measures are highly correlated with the percentage of students living in poverty, typically measured by the percentage of students participating in the federal free- and reduced-price lunch (FARM) program. Even though research has shown this percentage is a relatively poor proxy for student family income, especially when used in isolation,[xv] research has consistently shown a fairly strong correlation between the percentage of economically disadvantaged students as measured by participation in the FARM program and student outcome measures.

a. Cognitive Status Measures

As shown in Table 3, a number of school-level student characteristics were correlated with the percentages of students scoring proficient or advanced in the three subject areas. The strongest negative correlation was for the percentage of economically disadvantaged students while the strongest positive correlation was for the combined percentage of White or Asian students. In short, as the percentage of economically disadvantaged students in a school increases, the percentage of students scoring proficient or advanced in the school decreases. Alternatively, as the percentage of White/Asian students in a school increases, the percentage of students scoring proficient or advanced in the school also increases.

At the middle school and high school levels, the percentage of ELL students in a school was negatively correlated with the percentage of students scoring proficient or advanced. The correlation was moderately strong at the middle school level and strong at the high school level.

The percentage of students identified as gifted/talented was also positively correlated with the percentage of students scoring proficient or advanced, but not for all tests or at all school levels. The correlations were strongest and most consistent at the middle school and high school levels.

The percentage of students in special education (as defined by the percentage of students with an individualized education plan) was negatively correlated with the percentage of students scoring proficient or advanced bit not for all subject areas and not for all school levels. The correlations were strongest and most consistent at the high school level. There were no statistically significant correlations at the middle school level.

Interestingly, none of the correlations for the percentage of students scoring proficient or advanced on the IBSCA were nearly as strong as for the PSSA/Keystone results. Specifically, there was a weak, negative correlation with the percentage of economically disadvantaged students and a moderately positive correlation with the percentage of White/Asian students. These weaker correlations may be a result of the lack of variation in the characteristics of students and schools with IBSCA scores.

Table 3: Correlation Coefficients and Statistical Significance between Cognitive Status Measures and School Characteristics (2016)

SPP Indicator	Statistic	Percentage of Students						School Size
		econ disadv	White or Asian	Female	ELL	Gifted	Spec Ed	
LOWER ELEMENTARY SCHOOLS (Schools do not include grade 4)								
% Proficient/Advanced Mathematics	Correlation	-.662**	.492**	0.056	-0.144	.400**	-.189*	-0.067
	Stat Sig	0.000	0.000	0.506	0.085	0.000	0.023	0.425
% Proficient/Advanced Reading/ Eng Lang Arts	Correlation	-.674**	.511**	0.066	-0.142	.300**	-.204*	-0.077
	Stat Sig	0.000	0.000	0.429	0.090	0.000	0.014	0.359
% Proficient/Advanced Grade 3 Reading	Correlation	-.673**	.511**	.068	-.143	.303**	-.203*	-.074
	Stat Sig	.000	.000	.421	.088	.000	.014	.378
ELEMENTARY SCHOOLS (Schools include grade 4)								
% Proficient/Advanced Mathematics	Correlation	-.768**	.632**	0.035	-.307**	.281**	-.224**	-.052*
	Stat Sig	0.000	0.000	0.138	0.000	0.000	0.000	0.030
% Proficient/Advanced Reading/ Eng Lang Arts	Correlation	-.854**	.732**	.092**	-.411**	.434**	-.215**	0.038
	Stat Sig	0.000	0.000	0.000	0.000	0.000	0.000	0.116
% Proficient/Advanced Grade 3 Reading	Correlation	-.801**	.706**	.060*	-.382**	.434**	-.102**	-.092**
	Stat Sig	.000	.000	.034	.000	.000	.000	.001
% Proficient/Advanced Science	Correlation	-.664**	.679**	0.040	-.349**	.144**	-.177**	-.125**
	Stat Sig	0.000	0.000	0.093	0.000	0.000	0.000	0.000
MIDDLE SCHOOLS								
% Proficient/Advanced Mathematics	Correlation	-.862**	.854**	0.020	-.245**	.626**	-0.016	-0.009
	Stat Sig	0.000	0.000	0.744	0.000	0.000	0.800	0.887
% Proficient/Advanced Reading/ Eng Lang Arts	Correlation	-.846**	.849**	0.063	-.281**	.584**	-0.046	-0.025
	Stat Sig	0.000	0.000	0.300	0.000	0.000	0.450	0.682
% Proficient/Advanced Grade 3 Reading	Correlation	-.752**	.759**	.103	-.216**	.456**	-.078	-.088
	Stat Sig	.000	.000	.133	.001	.000	.255	.201
% Proficient/Advanced Science	Correlation	-.775**	.858**	0.123	-.264**	.496**	-0.013	-0.030
	Stat Sig	0.000	0.000	0.055	0.000	0.000	0.844	0.640
HIGH SCHOOLS								
% Proficient/Advanced Mathematics	Correlation	-.736**	.576**	.110**	-.355**	.418**	-.437**	.137**
	Stat Sig	0.000	0.000	0.004	0.000	0.000	0.000	0.000
% Proficient/Advanced Reading/ Eng Lang Arts	Correlation	-.762**	.614**	.182**	-.436**	.424**	-.493**	.109**
	Stat Sig	0.000	0.000	0.000	0.000	0.000	0.000	0.004
% Proficient/Advanced Grade 3 Reading	Correlation	-.539**	.523**	-.083	.012	-.148	.010	-.113
	Stat Sig	.001	.001	.629	.946	.388	.952	.511
% Proficient/Advanced Science	Correlation	-.832**	.757**	.145**	-.480**	.400**	-.470**	.117**
	Stat Sig	0.000	0.000	0.000	0.000	0.000	0.000	0.002
% College Ready on SAT or ACT	Correlation	-.807**	.608**	.080*	-.350**	.443**	-.420**	.191**
	Stat Sig	.000	.000	.041	.000	.000	.000	.000
% Proficient/Advanced IBSCA	Correlation	-.257**	.323**	.146**	-.195**	.138**	-.151**	.010
	Stat Sig	.000	.000	.003	.000	.005	.002	.838

Finally, there were strong correlations between the percentage of students scoring at the college ready level on either the SAT or ACT. Specifically, there was a strong negative correlation with the percentage of economically disadvantaged students and a strong positive correlation with the percentage of White/Asian students. Further, there were moderately strong negative correlations with the percentage of ELL and special education students and a moderately strong positive relationship with the percentage of gifted/talented students.

Thus, consistent with all other research in this area, the percentage of students scoring proficient or advanced on Pennsylvania state examinations were often strongly correlated with student background characteristics. In particular, the test results are most strongly correlated with the percentage of economically disadvantaged students and percentage of White/Asian students enrolled in a school. In addition, the college-readiness indicator was also strongly correlated with school-level student characteristics. While the IBSCA indicator was not strongly correlated with school-level student characteristics, the indicator was moderately correlated. These correlations strongly suggest that the indicators lack construct validity with respect to the assessment of the degree to which schools are effective.

Status measures—especially measures of student achievement--are correlated with school-level student characteristics for a number of reasons, many of which are outside the control of educators. For example, we know that many students living in poverty and some students of color come to school already substantially academically behind their more affluent peers and White students.[xvi] Thus, even if a school is successful in accelerating the learning of students coming to school academically behind, the use of status measures will likely identify the school as low-performing. We also know that students living in poverty suffer from "summer learning slide"[xvii] and typically experience higher levels of stress than their more affluent peers and these experiences have a negative impact on learning.[xviii] Status measures, then, simply cannot provide an accurate assessment of the effectiveness of a school or educators because status measures are heavily influenced by student demographics. Moreover, such students tend to suffer from "summer learning loss" and have less access to resources at home associated with school success. Finally, we know that for many students, living in poverty creates a high-level of stress which can impede the learning process.

b. Non-Cognitive Status Measures

Non-cognitive status measures include indicators such as attendance rates, student promotion rates, and cohort graduation rates. Research has consistently found that these measures are similar to status measures in that student and school characteristics heavily influence the rates. In Table 4 below, we present the relationships between the 2016 SPP non-cognitive measures (attendance rates, student promotion rates, and cohort graduation rates) and school-level student characteristics as well as school size. We employ the same shading strategy as in the previous analyses.

Across all school levels, there are statistically significant relationships between the non-cognitive indicators and school-level characteristics. These results suggest that these SPP non-cognitive measures do not accurately measure school effectiveness.. Again, this is particularly true at the high school level where the relationships are often moderate or strong and only one cell (graduation rates and school size) did not have a statistically significant relationship.

Table 4: Correlation Coefficients and Statistical Significance between Non-Cognitive Status Measures and School-Level Student Characteristics and School Size (2016)

SPP Indicator	Statistic	Percentage of Students						School Size
		econ disadv	White or Asian	Female	ELL	Gifted	Spec Ed	
LOWER ELEMENTARY SCHOOLS (Schools do not include grade 4)								
Attendance Rate	Correlation	-.668**	.483**	0.038	0.006	.317**	-.227**	-0.058
	Stat Sig	0.000	0.000	0.654	0.941	0.000	0.006	0.488
Promotion Rate	Correlation	-0.150	-0.048	0.070	.165*	0.135	-.229**	0.068
	Stat Sig	0.073	0.567	0.404	0.049	0.107	0.006	0.416
ELEMENTARY SCHOOLS (Schools include grade 4)								
Attendance Rate	Correlation	-.625**	.523**	.057*	-.209**	.289**	-.196**	0.000
	Stat Sig	0.000	0.000	0.018	0.000	0.000	0.000	0.999
Promotion Rate	Correlation	-.285**	.202**	0.015	0.024	.179**	-.060*	0.042
	Stat Sig	0.000	0.000	0.539	0.314	0.000	0.012	0.082
MIDDLE SCHOOLS								
Attendance Rate	Correlation	-.584**	.625**	.156*	-0.067	.420**	-0.013	-0.019
	Stat Sig	0.000	0.000	0.011	0.275	0.000	0.834	0.752
Promotion Rate	Correlation	-.125*	.258**	0.022	.124*	.237**	0.100	0.107
	Stat Sig	0.042	0.000	0.726	0.043	0.000	0.105	0.082
HIGH SCHOOLS								
Attendance Rate	Correlation	-.521**	.516**	.085*	-.420**	.250**	-.410**	.098*
	Stat Sig	0.000	0.000	0.026	0.000	0.000	0.000	0.010
Graduation Rate	Correlation	-.506**	.443**	.077*	-.421**	.282**	-.441**	-0.051
	Stat Sig	0.000	0.000	0.048	0.000	0.000	0.000	0.187

Source: PDE School Performance Profile Scores; Analysis: Dr. Ed Fuller

While the number and strength of the correlations between non-cognitive status measures and school characteristics were not as large or as strong as between the cognitive measures and school characteristics, the information presented in Table 4 above suggests that the non-cognitive status indicators in the SPP do not accurately capture school effectiveness. As such, the indicators lack the necessary construct validity to be used as measures of school effectiveness.

2. Growth Measures

In this section, we examine achievement growth measures. It is important to note that this is an incredibly complex topic and our discussion below is a simplified conceptual presentation. For further details, readers should consult the wealth of research studies that delve into this topic at a much deeper level than in this White Paper. Below, we focus on Pennsylvania's school growth measure, the Pennsylvania Value Added Assessment System (PVAAS), since the SPP uses the PVAAS results as part of the SPP calculation.

While status measures such as the percentage of students scoring proficient or advanced on the PSSA mathematics test capture student performance at a single point in time, growth measures are designed to assess the *change in student achievement over time*. On Pennsylvania, two different approaches to measuring growth are employed—one for PSSA assessments in reading and mathematics for grades four through eight and one for Keystone assessments and the PSSA science

and writing assessments. Two methodologies are employed because the data available to assess student growth differ for the two sets of assessments.

a. PSSA Reading and Mathematics Assessments in Grades 4 through 8

For these assessments and grade, test scores are available for multiple consecutive years. Because of the available data, PVAAS uses all of the testing history from prior years for a particular group of students to estimate the average achievement of the group. After the group of students actually takes the test during the current academic year, the new test scores from the most recent test administration are added to the prior scores to calculate a new estimate of the average achievement of the group of students. The new estimate of average achievement is compared to the estimate based on prior scores. The difference in the two estimates is the estimated academic growth which is then compared to the Pennsylvania standard for academic growth. Based on a statistical analysis, the academic growth of the group of students is classified as being in one of five groups. Before describing these five groups, there are some important points to consider.

First, PVAAS compares performance in two consecutive years. Ideally, the scores from the two years would be directly comparable. Some states have chosen to make the scores directly comparable by having a test that provides vertically aligned scores. With vertically aligned test scores, the scores from one grade level can be directly compared to the scores from another grade level. Suppose for example, that a state created a vertically aligned testing system in which the lowest possible score in the third grade was 200 and the highest possible score in the eighth grade was 1000. Further, let's assume student A achieved a score of 280 in the third grade and 320 in the fourth grade while student B achieved a score of 280 in the third grade but only a score of 300 in the fourth grade. We could conclude that, while both students made growth from the third grade to the fourth grade, student A made greater growth than student B. Pennsylvania, however, does not have a system with vertically aligned test scores. Thus, to compare test scores from one grade level to the next, the scores are converted to normal curve equivalent scores that range from 0 to 100. Because all scores are converted into a normal curve equivalent score, a NCE score of 40 in grade three and 50 in grade 4 would indicate greater than expected student growth while a NCE score of 40 in grade three and 40 in grade four would indicate expected student growth. While characterizing a NCE score of 40 in both years as meeting expected growth seems counter-intuitive, we have to understand that NCE scores are relative to all other test takers. Thus, we could think of this situation as a student scoring at the 4^{th} percentile of all students in the 3^{rd} grade and at the 4^{th} percentile of all students in the 4^{th} grade. Thus, the student maintained his position relative to all other test-takers. In this way, the student made the same growth as all other students. This is characterized as meeting expected growth.

Second, at the school level, the academic growth score has to be compared to some standard in order to determine of growth has been achieved. The PVAAS system does this by comparing the NCE scores for a group of students in one school to the same group of students in the same school relative to all other groups of students in all other schools. So, similar to the situation of the individual student above, a group of 4^{th} grade students that had a NCE score of 40 on the 3^{rd} grade test and then a 40 on the 4^{th} grade test would be characterized as meeting expected growth. Why? Because the achievement gain for the students maintained their achievement level relative to all other groups of students.

Third, statistical estimates of student test scores are only estimates, thus they involve error. In assessing academic growth, the error is called the Standard Error. The standard error allows us to place a confidence interval around the estimate of achievement. This is similar to the margin of error reported for political and other polls.

Using the above process, a group of students in a school is placed into one of the five following groups.

Group 1: The growth measure is greater than two standard errors below the average growth for the entire state; thus there is significant evidence for not meeting the standard for academic growth in Pennsylvania.

Group 2: The growth measure is greater than one, but equal to or less than two, standard errors below the average growth for the entire state; thus there is moderate evidence for not meeting the standard for academic growth in Pennsylvania.

Group 3: The growth measure is less than one standard error greater than the average growth for the entire state and less than one standard error below the average growth for the entire state; thus, there is evidence for meeting the standard for academic growth in Pennsylvania.

Group 4: The growth measure is equal to or greater than one, but less than two, standard errors above the average growth for the entire state; thus there is moderate evidence of exceeding the standard for academic growth in Pennsylvania.

Group 5: The growth measure is greater than two standard deviations above the average growth for the entire state; thus, there is significant evidence for exceeding the standard for academic growth in Pennsylvania.

b. Keystone Assessments, PSSA Science Assessment, and PSSA Writing Assessments

For these assessments, there are typically not test scores in consecutive years for students. Thus, PVAAS uses a different methodology to assess growth. To assess growth in these areas, PVAAS first uses all of the prior test scores for a group of students to predict the scores of the students on a Keystone exam, PSSA science exam, or the PSSA writing exam. Although the calculation is complex, a simplified version of the process by which a predicted score is calculated is that a student's predicted score is based on the observations of how students with identical testing histories scored on the test. In this way, there is an expected—or predicted—score for all students with the same scores on prior tests. When the student actually takes the test, then her or his score is compared to the predicted score that is based on how all other "identical" students scored on the test. Based on a statistical comparison of the actual scores for a group of students to the predicted score for the same group of students, PVAAS again places the group pf students into one of the five aforementioned performance groups.

c. Issues in Using School-Level Growth Measures in Accountability Systems

There are numerous important issues with using value-added growth measures in general, and models such as PVAAS, in particular in school accountability systems. The two most important questions about the use of school growth measures in accountability systems are:

(1) Do school growth models actually assess school effectiveness in improving student test scores? and,
(2) Do school-level growth measures accurately identify effective and ineffective schools?

Below, we address these two questions.

Do school growth models actually assess school effectiveness in improving student test scores?

Within the research field, there is general consensus that a simplified version of the statistical equation used to measure school-level student achievement growth and school effectiveness is captured in equation 1A below.

Equation 1A: Estimating School-Level Achievement Growth

Student Achievement Growth = prior scores + student characteristics + school inputs + school characteristics + school effectiveness

If we solve equation 1A to isolate school effectiveness, the result is equation 1B below.

Equation 1B: Estimating School Effectiveness

School effectiveness = student achievement growth - (prior scores + student characteristics + school inputs + school characteristics)

The PVAAS approach to measuring school-level student achievement growth and school effectiveness assumes that the use of prior student test scores removes the need to include school-level student characteristics, school inputs, and other school characteristics. Thus, the simplified version of the PVAAS approach is given in equation 2A.

Equation 2A: Estimating School-Level Achievement Growth Using PVAAS

Student Achievement Growth = prior scores + school effectiveness

The difference between the two approaches (Equations 1A and 2A) is the PVAAS approach assumes that the comparison of the scores of a group of students within a school to the prior achievement of that same group of students effectively captures the influence of student characteristics, school inputs, and other school characteristics. This assumption is based on the belief that student characteristics, school inputs, and other school characteristics outside the control of educators remain constant from one year to the next. The result of this assumption is that the statistical analysis used to measure student achievement growth does not need to include student characteristics, school inputs, or school characteristics. Thus, if we solve equation 2A for school effectiveness, we arrive at equation 2B below.

Equation 2B: Estimating School Effectiveness Using PVAAS

School effectiveness = Student Achievement Growth - prior scores

If, in fact, the prior test score history of groups of students within a school captures the influence of student characteristics, school inputs, and other school characteristics on student test score growth, then student characteristics, school inputs, and other school characteristics should not be correlated with school effectiveness. There would be little or no correlation because the inclusion of the prior scores of students in the statistical analysis would remove the influence of all of these factors on

student achievement growth. Up until 2016, the PVAAS vendor and PDE both maintained that the school-level PVAAS growth scores were, in fact, *not correlated* with any student characteristics, school characteristics, or school inputs.

However, reports by Research in Action and the Center for Evaluation and Education Policy Analysis (CEEPA) at Penn State demonstrated that school-level growth measures were, in fact, correlated with school level student demographics.

It is important to note that the SPP includes only converted school growth scores, not the actual school growth scores. For inclusion in the SPP, PDE converts school-level Average Growth Index (AGI) scores that range from roughly from around -25 to +25 into growth score values that range from 50 to 100. Schools with an AGI score of 3.0 or greater are assigned a value of 100 and schools with an AGI score of -2.92 or lower are assigned a value of 50. For schools with an AGI between -2.92 and +3.00, PDE assigned a growth score between 50 and 100, with higher AGI scores resulting in higher SPP growth scores.

The correlations between the truncated growth scores used in the calculation of the SPP and selected student demographics are presented in Table 5 for the 2016 SPP score.

Table 5: Correlation Coefficients and Statistical Significance between PVAAS SPP Growth Measures and School-Level Characteristics School Size (2016)

SPP PVAAS Student Growth Measure	Statistic	Percentage of Students						School Size
		econ disadv	White or Asian	Female	ELL	Gifted	Spec Ed	
ELEMENTARY SCHOOLS								
SPP Math	Correlation	-.044	.054*	-.015	.002	.023	.027	.047
	Stat Sig	.101	.044	.583	.937	.390	.322	.081
SPP Reading	Correlation	-.129**	.114**	-.024	.024	.062*	-.009	.067*
	Stat Sig	.000	.000	.376	.380	.021	.739	.012
SPP Science	Correlation	-.294**	.399**	-.001	-.205**	.123**	.058*	-.031
	Stat Sig	.000	.000	.982	.000	.000	.035	.254
MIDDLE SCHOOLS								
SPP Math	Correlation	-.124**	.126**	.030	-.027	.037	-.129**	-.059
	Stat Sig	.002	.001	.441	.491	.344	.001	.131
SPP Reading	Correlation	.066	.007	.054	.045	-.156**	-.108**	-.126**
	Stat Sig	.094	.866	.169	.249	.000	.006	.001
SPP Science	Correlation	-.403**	.495**	.015	-.253**	.218**	-.063	-.024
	Stat Sig	.000	.000	.707	.000	.000	.108	.537
HIGH SCHOOLS								
SPP Math	Correlation	-.394**	.329**	.022	-.228**	.174**	-.189**	.063
	Stat Sig	.000	.000	.574	.000	.000	.000	.102
SPP Reading	Correlation	-.259**	.179**	.068	-.166**	.096*	-.188**	.028
	Stat Sig	.000	.000	.079	.000	.013	.000	.476
SPP Science	Correlation	-.484**	.365**	.010	-.228**	.184**	-.219**	.159**
	Stat Sig	.000	.000	.790	.000	.000	.000	.000

Source: PDE School Performance Profile Scores; Analysis: Dr. Ed Fuller

Clearly, the results for the PVAAS statistical approach employed for the science tests in elementary school and middle school are, in fact, correlated with various school-level

characteristics and some of these correlations are fairly strong. Further, the PVAAS statistical approach to calculate growth for the Keystone examinations are also correlated with student characteristics, particularly in 2016 in which all three Keystone growth results were negatively correlated with the percentage of economically disadvantaged students enrolled in the school. Thus, for the PVAAS approach used to estimate school-level growth when there are not test results from two consecutive grades clearly yields results that are correlated with school-level student characteristics.

Thus, the PVAAS growth model for science at the elementary and middle school levels as well as for the Keystone results at the high school level do not remove the influence of school-level student characteristics on the estimate of student growth. Interestingly, in response to the revelations of correlations between PVAAS growth scores and school-level student characteristics, PDE wrote the following in their *Response to PVAAS Misconceptions* booklet:

> There are a few subjects and grades (namely, PSSA Science and the Keystones) where there is a small or moderate relationship between growth and students from certain subgroups. In interpreting these results, it must be emphasized that, *at the individual student-level and taking into account a student's prior testing history, characteristics like the socioeconomic status of that student does not have a relationship to the student's ability to show growth*. Data from other states indicates that there is typically no relationship between growth measures and demographics, even in the end-of-course or science assessments. With that in mind, the results of this reporting in PSSA Science and the Keystones is an opportunity to re-assess whether the standards-aligned system is fully implemented in all classrooms, schools and districts and whether there are additional needs and supports for certain student populations, schools and districts. PDE will continue to monitor results. [xix]

Surprisingly, this statement assumes without any evidence that schools with greater proportions of economically disadvantaged students may simply have not yet aligned their curriculum and pedagogy with state standards and the state assessments. While this may be true, there is no evidence that this claim is, in fact, accurate. Moreover, there is no evidence that this mis-alignment is the cause of the correlations between student demographics and the aforementioned growth results. Equally plausible explanations are that (1) the growth methods employed for elementary science examinations and the Keystone examinations are not capable of removing the influence of student personal characteristics or that (2) school-level student characteristic are a proxy for other factors that are influencing student scores such as access to effective teachers or school inputs such as school funding levels. Indeed, the Pennsylvania school finance system is one of the most inequitable in the country and that Pennsylvania schools serving high proportions of economically disadvantaged students and racial/ethnic minority students are disproportionately located in school districts that are underfunded. This could certainly be the cause of the aforementioned correlations.

Unfortunately, despite the fact that PDE has been well aware of this situation for nearly one year, there are still no publicly available reports that investigate this issue in any greater depth. Further, PDE has not made data available to researchers that would allow researchers to explore this issue. At the very least, PDE should have released a statement with the SPP scores that discusses this issue and cautions that the methods employed for estimating growth for elementary science examinations and Keystone examinations may not be accurate.

d. Do school growth models provide information that allows schools to be accurately rank-ordered?

Another issue related to growth scores is the degree to which growth model statistical estimates allow for the accurate "stack" ranking of schools. By stack ranking, we mean the ranking of schools based on their growth measure. To some degree, PDE's decision to transform the AGI scores into the SPP growth measure scores that have a base of 50 and a cap of 100 reduces the impact of the statistical estimates on the ranking of schools by grouping schools with very low and very high AGI scores. Indeed, as shown below in Table 5, between 45% and 16% of schools are included in one of these two categories.

Table 5: Number and Percentage of Schools Included in the Ranges of Truncated Growth Scores in the SPP by School Level

Growth Score	Mathematics		Reading/ELA		Science	
	Number	Percent	Number	Percent	Number	Percent
ELEMENTARY SCHOOLS						
50	215	12.2	203	11.6	239	13.7
51 to 99	1242	70.7	1348	76.7	1174	67.4
100	299	17.0	206	11.7	329	18.9
All Schools	1756	100.0	1757	100.0	1742	100.0
MIDDLE SCHOOLS						
50	46	17.2	26	9.7	24	9.9
51 to 99	177	66.0	205	76.5	204	84.0
100	45	16.8	37	13.8	15	6.2
All Schools	268	100.0	268	100.0	243	100.0
HIGH SCHOOLS						
50	170	24.8	102	14.9	121	17.7
51 to 99	380	55.5	479	69.9	416	60.7
100	135	19.7	104	15.2	148	21.6
All Schools	685	100.0	685	100.0	685	100.0

However, this means that between 84% and 55% of schools are "stack" ranked with a growth measure score in the SPP. Even if we assume that the PVAAS statistical model removes the influence of school-level student characteristics and other school characteristics on the growth score, the exclusion of such factors can influence the stack rankings of schools.[xx] For example, based on their investigation into the effects of including or excluding student background characteristics when ranking schools based on growth scores, Ehlert and his colleagues (p.26) state the following:

> . . . the decision about whether to control for student covariates and schooling environments, and how to control for this information, influences which types of schools and teachers are identified as top and bottom performers. Models that are less aggressive in controlling for student characteristics and schooling environments systematically identify schools and teachers that serve more advantaged students as providing the most value-added, and correspondingly, schools and teachers that serve more disadvantaged students as providing the least. Given recent arguments in favor of using equally circumstanced comparisons in education evaluations (Barlevy

> and Neal 2012; Ehlert et al. 2013)—that is, comparisons between schools and teachers that serve similar student populations—this is an important consideration for state and local education agencies that are exploring the use of value-added models as a part of their accountability systems.[xxi]

In other words, controlling for student- and school- characteristics and the methods employed to control for the influence of these factors that are outside of the control of educators can alter the schools identified as the bottom- and top- performing schools. This is an extremely important point given that ESSA mandates that states identify the lowest performing 5% of schools and create interventions for such schools. Given that including or excluding student- and school-characteristics and the methods for including these factors in a growth model could possibly change the schools identified as being in the lowest performing 5% of schools, PDE should be transparent about which schools are identified as the lowest performing 5% of schools under different methodological approaches. In the same report, PDE should present the stack rankings of all schools under different growth model approaches so that educators and policymakers can make informed decisions about policies and the designation of schools as low-performing.

e. Other Issues Related to the Calculation of Growth Scores

There are several other important issues related to the calculation of school-level growth scores. These issues include: (1) the characteristics and properties of the test used in a state; (2) assumptions about student characteristics; (3) the type of growth model employed; (4) the use of student background variables in a growth model within a particular type of approach

i. Test Characteristics

Research has generally found that different tests will yield different growth score results for both students and school—even when the exact same growth model is used. This occurs, in part, because different tests measure different sets of knowledge and skills. Further, use of tests by states assumes the tests have the appropriate psychometric properties. Research suggests the degree to which state tests meet the correct psychometric properties varies across states.

ii. Assumptions about Student Characteristics

In their description of the growth model approach used to calculate PVAAS, the vendor claims that student characteristics remain "consistent" over time. Yet, this claim is not substantiated by any publicly available analyses and PDE has repeatedly refused to release data to independent researchers to test this claim. There is evidence, in fact, that student characteristics do change over time. For example, during the recession, a substantial number of students moved from not living in poverty to living in poverty. Our own analyses of this phenomenon using student-level data from Texas suggests that about 10% of students not identified as participating in the FARM program in 8th grade were, in fact, identified as participating in 9th grade in the following year. The opposite movement occurred as well—about 12% of students identified as participating in the FARM program in the 8th grade were not identified as participating in the FARM program in the 9th grade during the following year. This second finding underscores an important research finding—students are less likely to participate in the FARM program as they get older, regardless of whether they remain eligible to actually participate in the program.

Further, with respect to participation in special education and English Language Learner programs, we know that students transition into and out of such programs every year. Indeed, using Pennsylvania student-level data from the 2010-11 and 2011-12 academic years, we found that greater than 2% of students not participating in special education in 2010-11 did participate in special education in 2011-12. In addition, 6% of students who participated in special education in 2010-11 did not participate in special education in 2011-12. More startlingly is the finding that about 6,000 students who were enrolled in multiple schools during the 2011-12 school year were identified as participating in special education in one school, but not in the other school. Thus, the assumption that student characteristics remain consistent over time is demonstrably false.

Moreover, there are also unobserved characteristics of students that affect student performance such as physical and mental health issues, stress at home caused by life events, and other issues that could never be captured by any data collection system. Over time, a value-added model will capture these effects through the use of prior scores. However, the value-added calculation will not capture the effects of these changes in the year in which they occur. For example, suppose a student lives with one parent and that parent loses her job. While the student did not live in poverty the prior year, the student now lives in poverty and suffers all the stressors from living in poverty. The prediction made by the value-added model would be based on prior scores for the student during which time the student did not experience the stressors associated with living in poverty. Thus, the VAM might over-predict the growth expected by the student because the prediction could simply not account for the impact of the change in the student's socio-economic status on her/his learning trajectory. While the aggregation of student scores to the school level may substantially reduce the impact of such situations on a school-level growth score, the effect would be more pronounced in schools with smaller student enrollments than in schools with larger student enrollments.

Even if the observed characteristics of a student such as participation in the FARM program, special education, and bilingual education remain consistent over time, we know that these measures are often only poor proxies for information about a particular student. For example, there is ample evidence that participation in the federal FARM program is a fairly inaccurate proxy measure about a student's socio-economic situation. One reason for this inaccuracy is that the measure is largely binary—a student either participates or does not. Thus, a student whose family is $1 above the threshold to be eligible for the FARM program is treated identically as a student whose family makes hundreds of thousands of dollars a year. Moreover, if a student's family situation changes from having a family income of $500,000 to $75,000 per year, the student's FARM participation measure will not change despite the substantial change in the student's family situation. Even more importantly, research has shown that poverty is often not the cause of the poor academic performance of a student, but the stressors in a student's life that often occur because of living in poverty are what influence a student's ability to perform academically. In fact, students across the socio-economic spectrum can experience life stressors (lack of food, mental or physical abuse, mental health issues, etc) that negatively impact student performance and none of these stressors are accurately captured by any data system nor could any data system ever be expected to capture such information.

iii. Type of Growth Model Employed by a State

There are a number of different ways to calculate growth models and the decisions about which growth models to use can substantially alter the outcomes of the value-added calculations. Thus, schools that appear to be exceeding expected growth under one value-added calculation might appear to only be meeting expected growth using a different model to

calculate the growth score. When originally choosing to hire a vendor that employed a specific growth model, PDE should have made public an analysis of how the stack rankings of schools would change under the use of different growth models. In fact, if the state would release data to researchers, such an analysis could be conducted now as a means to inform parents, educators, taxpayers, and policymakers about how different approaches to measuring school-level growth affects the rankings of particular schools.

iv. Issues with Grade Levels in Growth Measures

An additional issue that has not been addressed by PDE is that research suggests VAMs in general[xxii] and the EVAAS VAM approach in particular[xxiii] may function differently across grade levels or grade spans. If, in fact, PVAAS functions differently across grade levels, then schools could be advantaged or disadvantaged simply by having a specific type of grade configuration for their school. This would violate a number of different types of validity issues, including construct validity.

Because we did not have access to individual student data, we could not thoroughly investigate this issue. Moreover, we could not find any technical report by either PDE or PVAAS that would shed light on this issue. Clearly, this is information that should be made available to the public as a strategy to increase confidence in the system and, consequently, improve the face validity of PVAAS and the SPP.

3. Achievement Gap Measures

One primary focus of NCLB was the reporting of student achievement results by student sub-populations as a means to encourage schools to close the achievement gap between various sub-populations of students. There is, however, no evidence that the reporting of achievement gaps or the inclusion of accountability gap measures in accountability systems influences actual changes in the achievement gap.[xxiv] Indeed, Harris and Herrington find that the only accountability measures associated with closing the achievement gap are those accountability measures that increase access to and equitable distribution of either (a) resources or (b) quality curricula.[xxv] Thus, there is no extant evidence that achievement gap measures have any predictive validity.

Moreover, the SPP incorrectly calculates the achievement gap because the calculation is based on percentages of students scoring proficient or advanced.[xxvi] While we do not review the underlying reasons for why the use of percentages invalidates the SPP achievement gap measure, research is clear that any achievement gap measure based on percentages of students scoring at a particular level does not have construct validity or any other form of validity. Indeed, such achievement gap measures simply do not allow for any valid inferences to be made from the data.

4. Overall SPP Score

The SPP score is calculated by summing the points earned by each school for each indicator and then dividing the total points earned by the total points possible. Each indicator is assigned a weight and the indicator score is multiplied by the weight to arrive at a point value for that particular indicator. In this section, we examine the degree to which the overall SPP measures school-level characteristics or school effectiveness apart from school characteristics.

As shown in Table 6, there is a strong negative relationship between the percentage of economically disadvantaged students in a school and the SPP scores at each of the four school levels such that, as the percentage of economically disadvantage students increases, the SPP score

decreases. Overall, the percentage of economically disadvantaged students in a school explains between 25% and 35% of the variance (dispersion) of scores.

Table 6: Correlation Coefficients and Statistical Significance between SPP Scores and School-Level Student Characteristics and School Size (2016)

School Level	Statistic	Percentage of Students						School Size
		econ disadv	White or Asian	Female	ELL	Gifted	Spec Ed	
Lower Elementary	Correlation	-.515**	.487**	.032	-.107	.188*	-.115	-.074
	Stat Sig	.000	.000	.729	.247	.042	.214	.427
	N	118	118	118	118	118	118	118
Upper Elementary	Correlation	-.658**	.659**	.058*	-.336**	.216**	-.099**	-.096**
	Stat Sig	.000	.000	.049	.000	.000	.001	.001
	N	1149	1149	1149	1149	1149	1144	1149
Middle School	Correlation	-.706**	.760**	.101	-.166*	.442**	-.016	-.067
	Stat Sig	.000	.000	.122	.011	.000	.801	.303
	N	236	236	236	236	236	236	236
High School	Correlation	-.630**	.548**	.266**	-.371**	.385**	-.530**	-.005
	Stat Sig	.000	.000	.000	.000	.000	.000	.946
	N	189	189	189	189	189	189	189

While we would expect the percentage of economically disadvantaged students in a school to be somewhat negatively related to school effectiveness because of greater teacher/principal turnover, greater percentages of inexperienced teachers/principals, and relatively less access to revenue, we would not expect the relationship to be nearly as strong as portrayed in Table 6.

The strength of the relationship is due to nearly one-half of the available SPP points coming from status indicators (percent proficient/advanced on Keystone Exams; percent competent/advanced on Industry Standards-Based Competency Assessments; and, percent of students meeting SAT/ACT college-readiness standards) and other indicators (cohort graduation rate; attendance rate; percent participating in Advanced Placement/International Baccalaureate Diploma, or College Credit; and PSAT/Plan participation) that are highly correlated with school-level student demographics. Moreover, as shown previously, the growth indicators are also correlated with the percentage of economically disadvantaged students as are some of the gap closure indicators.

When we include other school-level demographics and school size in the analysis, 61% of the variance in the SPP scores is explained by factors *not under the control of educators*. Unfortunately, without access to additional data and more detailed data, we cannot accurately determine what, in fact, the SPP is measuring. The SPP could be measuring the influence of student characteristics—especially poverty—on student achievement. The SPP could also be measuring differential access to fiscal and human resources that influence student outcomes. While we cannot conclusively identify what, in fact, the SPP is measuring, we are quite confidant in asserting that the SPP is *not* an accurate measure of school effectiveness. . In short, the SPP is an inaccurate measure of school effectiveness. This does not mean the SPP does not include useful information. *Rather, the SPP should not be used as a measure of school effectiveness.*

Using the 2014 SPP results, we find very similar results for elementary schools and middle schools. In fact, at least at the high school level, the 2014 SPP scores are even more strongly associated with student characteristics than the 2015 SPP scores. The different weights in the 2015 system and the inclusion of more achievement gap measures slightly reduced the correlation.

D. Potential Consequences of Using the Current Configuration of the SPP

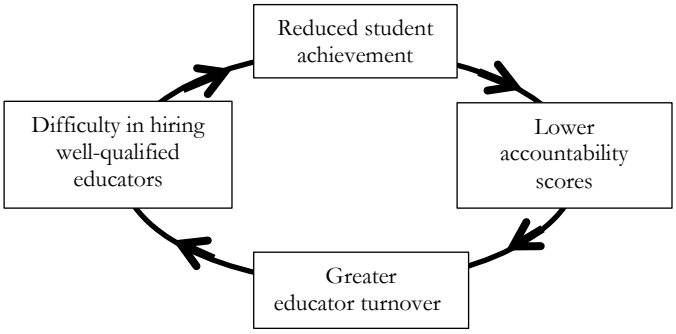
Why does it matter if the SPP score is highly correlated with school-level student demographics? There are, in fact, fairly substantial negative consequences of using the current SPP system or any system that is strongly correlated with school-level student characteristics.[xxvii]

First, providing inaccurate information about school effectiveness to educators sends the wrong signals about performance, thus subverting the primary mechanism through which accountability systems drive improvement. For example, inaccurate accountability systems often signal to schools enrolling large percentages of White and affluent students that their policies, procedures, and strategies are effective, thus sometimes leading to complacency.[xxviii] At the other end of the continuum, inaccurate accountability systems signal to schools enrolling large percentages of students living in poverty and students of color that their policies, procedures, and strategies are ineffective, thus often leading to the rapid adoption of new policies.[xxix] This rapid re-tooling of approaches greatly reduces the odds such schools will improve over time.[xxx]

Second, the SPP is currently used as one component of teacher and principal evaluation systems, thus creating an unfair system of evaluation. In reviewing teacher and principal accountability systems, Helen Ladd notes, "An approach would be unfair if it attempted to hold the teachers and a principal of a school accountable for factors beyond their control" (p. 391).[xxxi] This is precisely what we have shown above: the inclusion of the SPP score in teacher and principal evaluation holds educators accountable for factors (student demographics and school size) outside of their control.

Third, the inclusion of an unfair accountability system score in teacher and principal evaluations create nearly insurmountable barriers to the recruitment and retention of effective teachers and leaders.[xxxii] The increased difficulty in hiring and the constant churn of educators creates a vicious downward spiral of schools serving students living in poverty and students of color as displayed in Figure 5 below.

Figure 5: Downward Spiral of Lower-Performing Schools

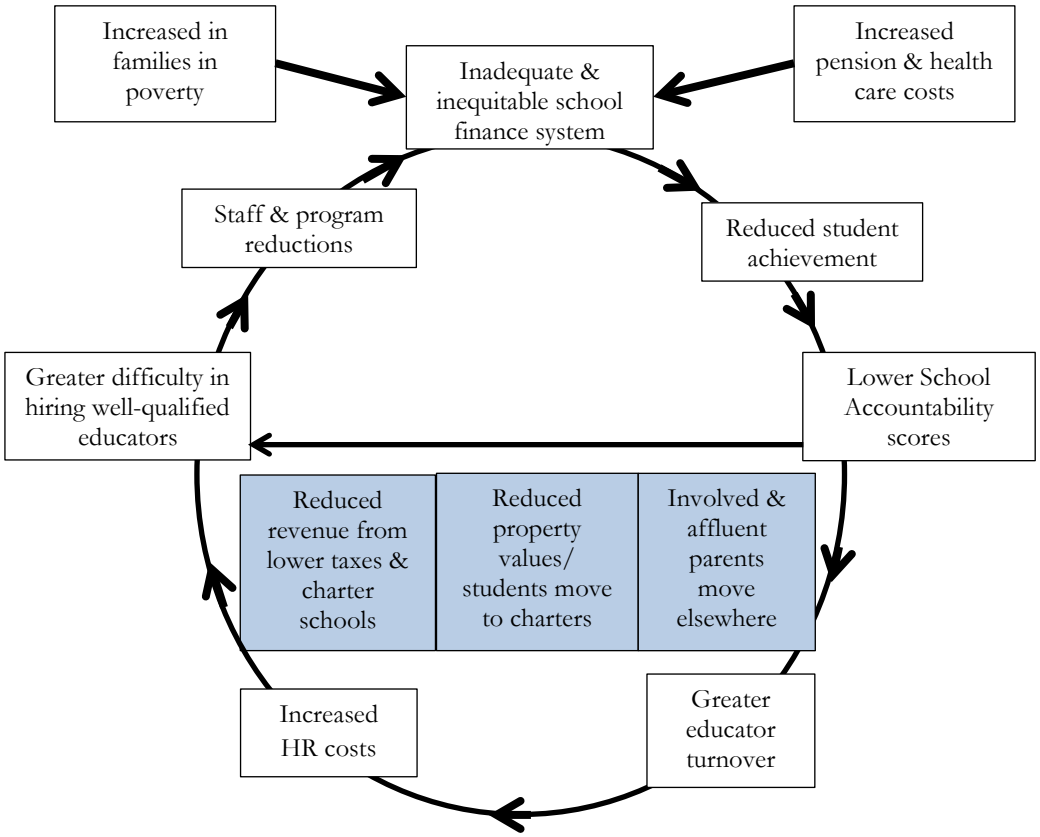


Third, research has consistently found that decreases in school and district accountability ratings leads to: (a) more involved families choosing to send their children to other schools; and, (2) a decline in housing prices.[xxxiii] The increase in more involved families enrolling their children in

other schools often leads to a further decline in accountability ratings and deepens the cyclein our more robust description of what we term the "PA Public School Death Spiral" shown in Figure 6.. The decline in housing prices depresses the funding generated by local property taxes. Given that schools in the Commonwealth rely more heavily on property taxes for school funding than in many other states and are limited by Act 1 in the amount of the tax increases they can propose, the decrease in housing prices typically diminishes the revenue available to a district. Further exacerbating this problem is that way in which Pennsylvania funds charter schools—particularly with respect to special education students. All of these factors interact to create a situation in which lower-performing schools have increasingly less money to spend on a student population that is increasingly more difficult to educate. Thus, the current configuration of the SPP has serious detrimental effects in a number of ways and, most importantly, seriously disadvantages schools serving high-poverty communities and communities of color.

While schools serving large percentages of students living in poverty and students of color were likely ensnared in this cycle prior to the public release of test scores and school accountability systems, the adoption of inaccurate school accountability metrics has simply hastened and further entrenched this cycle. The odds of a low-performing school escaping this cycle without changes in state policies, including those governing school accountability systems and resource allocation efforts, are minimal. While some schools are able to escape this cycle, the vast majority of schools cannot without adopting policies and procedures that do not meet the needs of all children such as "no excuses" discipline policies and the counseling of families with special needs children to look at other school options. We will discuss this issue further in our discussion about the identification of low-performing schools.

Figure 6: Factors Involved in the Pennsylvania Public School Death Spiral



II. PROPOSAL FOR A NEW SCHOOL PERFORMANCE PROFILE AND SCHOOL ACCOUNTABILITY SYSTEM

As mentioned previously, PDE has already recognized that there are a number of issues with the SPP that need to be changed. In fact, PDE has gathered substantial input from educators and others from around the state and should be applauded for their efforts in this area. From our perspective, there are two major problems with the current configuration of the SPP: first, the system does not treat all schools fairly because of the correlation between the overall SPP scores and factors outside the control of the school; and, second, the system is too narrowly focused on test-based measures such that the SPP score provides information only about a narrow slice of the overall purposes of education

With respect to system fairness, Marion (2016) contends that, "One of the key tenets of accountability design is that the results of applying the accountability rules should not privilege or reward schools based on the demographic characteristics of the school" (p. 2).[2] We subscribe to this perspective on accountability systems and, thus, focus our suggestions on strategies the Commonwealth might employ to create a SPP system that is less a measure of school-level student characteristics and more a measure of school effectiveness.

This is a difficult endeavor given the constraints that ESSA places on states. Despite the rhetoric about ESSA devolving substantial authority to states about designing new accountability systems, ESSA does place restrictions upon states in their attempts to design completely new accountability systems. In particular, the focus on test scores as a substantial component of accountability systems and the requirement that states include test-based achievement outcomes for major student sub-groups (economically disadvantaged, White, Black, Hispanic, special education, English Language Learner) as part of the school accountability system forces states to heavily weight indicators that stem directly from state standardized tests. This essentially guarantees that school accountability systems will be correlated with school-level student demographics. In fact, based on the most likely components of the SPP, it would be quite difficult, if not impossible, to devise a system that is not correlated to some degree with school-level student demographics.

With respect to the narrow focus of the current configuration of the SPP, there is widespread agreement that the federal initiatives in this area (No Child Left Behind, No Child Left Behind Waivers, and Race to the Top) forced states to construct narrowly defined school accountability systems. Indeed, even federal policymakers recognized that this was an issue and mandated that states include accountability measures that do not stem directly from test scores.

The problem that states face—including Pennsylvania—is selecting non-test measures that accurately capture important educational outcomes and, simultaneously, are not influenced by student characteristics.

1. Recommendations on SPP and SPP Indicators

In this section, we provide a number of recommendations for improving the SPP. PDE has not released any information on metrics that are under considerations, thus we could not review them as part of this paper.

a) Fund an Independent Study of the School Performance Profile System

The Commonwealth should post a request for proposals that invites researchers to submit applications to conduct an in-depth study the SPP system. The funding for such a study should be

[2] Marion, S. (2016). *Considerations for State leaders I the Design of School Accountability Systems under the Every School Succeeds Act.* Dover, NH. National Center for the Improvement of Educational Assessment.

substantial given that such a study would need to include both quantitative and qualitative measures. To facilitate a high-quality study, PDE should make available any and all data requested by the researchers.

The study should focus on determining the degree to which the SPP system yields information that allows for reliable and valid inferences to be made about schools. Included in this analysis should be a comparison of multiple SPP approaches and how the scores and rankings of schools change as a result of changes in the SPP. For example, the study should document how changes in the weighting of the different indicators change the scores and rankings of schools.

b) Weight Growth and Adjusted Non-Cognitive Indicators as Heavily as Possible

As we have shown above, school accountability systems attempt to meet multiple purposes. At this point in time, we have not seen any state school accountability systems that adequately meet both purposes. If a state places a relatively low weight on proficiency measures and greater weights on growth measures and other measures to reduce the influence of school-level student characteristics, then it may be possible to satisfy both purposes. However, given that ESSA requires that (a) states must provide an overall score or grade, (b) states must count achievement indicators as a substantial proportion of the overall score/grade, and (c) states must include major student sub-groups in the reporting of achievement measures, we think it is rather unlikely that Pennsylvania and other states will be able to construct an accountability system that is not correlated with school-level student characteristics. The inability of states to construct school accountability systems in such a way certainly diminishes the validity of the systems and, thus, ultimately reduces the effectiveness of such systems.

One possible configuration of the SPP is displayed in Table 7. If certain conditions are met, then approximately 74% and 72% of the SPP score would not be correlated with school-level student characteristics for elementary/middle schools and high schools, respectively. These conditions would include the following:

- Growth measures, including English Language Proficiency progress, are not correlated with school-level student characteristics;
- Locally Selected Indicators are not correlated with school-level student characteristics.

Whether such conditions can be met remains unseen. Unfortunately, we do not have access to sufficient data to test our proposal. We do believe that such an approach is worth exploring and, as such, recommend that the Commonwealth enlist researchers to study the feasibility of crafting such a system of indicators. We also strongly urge PDE and the Legislature to push the envelope as much as possible and attempt to create a school accountability system that meets both purposes.

Extreme caution must be undertaken when adjusting scores using only the observable school-level student characteristics. By observable characteristics, we mean characteristics on which we collect data such as participation in the FARM program, student race/ethnicity, student gender, and student participation in special education, gifted education, and English Language Learner programs. There, are, however, unobservable characteristics of students such as parental level of education, parental level of support, parental involvement in education, ambition, resiliency, and effort to name just a few. If such characteristics are randomly sorted across schools, then we do not have to worry about the issue of unobservable student characteristics. If, however, such characteristics are not sorted randomly across schools, then the adjusted measures may be inaccurate. There is, in fact, some evidence that students with some of these characteristics sort themselves into charter schools and magnet schools. This is why the higher quality studies of

charter school achievement attempt to compare students that entered a lottery and entered a charter school to students that entered the lottery but did not enter a charter school. Such a method partially controls for the unobservable characteristics of students (and their parents/guardians). This issue should be thoroughly explored before the Commonwealth adopts a strategy of adjusting specific indicators within the SPP. We want to underscore, however, that the real issue is not whether the Commonwealth should attempt to adjust the indicators for factors outside the control of schools, but the degree to which the Commonwealth—with the assistance of researchers—can adequately adjust the indicators to construct a fair and accurate school accountability system.

Table 7: Potential SPP Configuration
to Reduce Correlation with School-Level Student Characteristics

SPP Indicator	El/MS	HS
Indicators of Academic Achievement	**19**	**21**
Mathematics/Algebra I - Percent Proficient or Advanced on PSSA/Keystone	**8**	**8**
All Students	2	2
Required Student Sub-Population Groups	6	6
ELA/Literature - Percent Proficient or Advanced on PSSA/Keystone	**8**	**8**
All Students	2	2
Required Student Sub-Population Groups	6	6
Science/Biology - Percent Proficient or Advanced on PSSA/Keystone	**3**	**3**
All Students	2	2
Required Student Sub-Population Groups	1	1
Industry Standards-Based Competency Assessments - % Competent or Above	**0**	**2**
All Students	0	2
Indicators of Academic Growth/PVAAS	**48**	**48**
Mathematics/Algebra I - Meeting Annual Academic Growth Expectations	**16**	**16**
All Students	8	8
Econ Disadv Students	8	8
ELA/Literature - Meeting Annual Academic Growth Expectations	**16**	**16**
All Students	8	8
Econ Disadv Students	8	8
Science/Biology - Meeting Annual Academic Growth Expectations	**16**	**16**
All Students	8	8
Econ Disadv Students	8	8
Progress Toward English Language Proficiency	**7**	**5**
Other Academic Indicators	**16**	**16**
Adjusted Promotion Rate	4	0
Adjusted Attendance Rate	3	2
Participation rate in employment exploration	3	2
Cohort Graduation Rate	0	2
Adjusted Graduation Rate	0	3
Adjusted % of Graduates Entering Higher Education or Employment	0	3
Student discipline rates	2	2
Use of school climate survey	4	2
Locally Selected Indicators	**10**	**10**
TOTAL	**100**	**100**

There are four major changes in this proposal relative to the current SPP configuration. First, there are no achievement gap measures in this proposal. Second, the ratio of achievement growth measures to achievement status measures is increased substantially. Indeed, the ratio in the current SPP is 1:1 while the ratio in the proposed system is greater than 2:1. Assuming PDE includes student background characteristics in the calculation of PVAAS scores, this shift would dramatically alter the degree to which the overall SPP score was correlated with school-level student characteristics. Indeed, the correlations should diminish rather substantially. Third, the number of non-cognitive status measures is increased and the overall weight (or influence) of these indicators is also increased. This change would accomplish two purposes—the reduction of indicators based on test scores and the reduction of the correlation between the overall SPP score and school-level student characteristics. Fourth, a new section worth 10 points is included that allows districts to choose additional indicators. The indicators would be proposed to PDE and PDE, with the assistance of researchers, would approve the indicators for use in the new SPP. These changes are discussed below or elsewhere in this report.

c) Include Additional Indicators

Even if PDE does choose to apply a heavier weight to the school growth measure and adjusts the current non-cognitive measures, the overall SPP score would still stem primarily from the points directly associated with test scores. Given the growing consensus that the myriad purposes of education cannot be measured based almost exclusively on indicators derived from test scores, the Commonwealth should expand the number of non-cognitive outcomes. However, in doing so, the state should engage researchers on the various indicators that could be adopted and how such indicators could be statistically adjusted to remove the influence of factors outside the control of educators.

d) Include Student Characteristics in the PVAAS Calculations for the SPP

While PVAAS is constructed to not be correlated with school-level student characteristics, we have demonstrated that, in fact, PVAAS is correlated with some school-level student characteristics for some tests at some school levels. This is true regardless of whether we use the full range of PVAAS scores (AGI scores) or the truncated growth scores that are used in the SPP. Thus, PDE should ensure the vendor calculates a growth measure that removes the influence of school-level student characteristics.

Moreover, the vendor should submit annual technical reports that provide the foundational information used in the calculation of the value-added measures. For example, these technical reports should provide the value added scores for each grade level and subject area (e.g., 4th grade mathematics), the number of test-takers in each grade level and for each subject area, and a description of the method used to arrive at the overall school value-added score (e.g., weighted versus an unweighted average). This technical report should also include the average value-added score by prior year score to document that the value-added score is not associated with a student's prior score and that high-performing students can, in fact, achieve greater than expected growth to the same degree as lower preforming students.

Further, PDE should report a student mobility indicator for each school and the PVAAS vendor should include this indicator in the PVAAS methodology. This is an important factor that influences school-level outcomes that is generally considered to be largely out of the control of educators. As such, the PVAAS model should remove the influence of student mobility on student growth measures.

Finally, and perhaps most importantly, PDE and the value-added vendor should make clear to the public and policymakers how changes in the value-added model change the stack rankings of

schools. Research has shown that the model employed to calculate the value-added scores can radically change the stack rankings of schools. Yet, this information has never been made public in the Commonwealth. Given the research evidence that the inclusion or exclusion of student demographics alters the stack ranking of schools when using adjusted status measures or growth measures, PDE should either (a) instruct the vendor to include student demographics in the PVAAS calculations or (b) provide evidence that the exclusion of student demographics does *not* alter the stack rankings of schools.

e) Explore the Adjustment of Indicators for School-Level Student Characteristics

For measures other than proficiency or growth measures (such as attendance rates, graduation rates, percentage of students enrolled in advanced classes, percentage of students taking SAT/ACT/AP/IB tests, percentage of students achieving a particular score on SAT/ACT/AP/IB tests, enrollment in post-secondary education, and employment), PDE should use regression analysis to remove the influence of school-level student characteristics to the greatest degree possible. If ESSA allows for states to use such adjusted measures rather than the simple percentages, then the Commonwealth should use the adjusted measures to the greatest degree possible. The use of adjusted indicators will (1) provide more accurate indicators of school effectiveness by removing the influence of school-level student characteristics on the indicators and (2) decrease the correlation between the overall SPP score and school-level student characteristics.

Not only are current SPP indicators correlated with school-level student characteristics, proposed additional SPP measures would also likely be correlated with school-level student characteristics. For example, in Table 8, we examine the likely correlations between school-level student characteristics and proposed new or modified SPP indicators as suggested by PSBA.[xxxiv] If left unadjusted through statistical procedures, the proposed indicators would simply reinforce the current SPP's correlation with school-level student characteristics.

Table 8: Potential Correlations between Proposed Additions and Modifications to SPP Indicators and School-Level Student Characteristics

Potential Indicator	Percentage of Students					School Size
	Econ Disadv	Students of Color	ELL	Gifted	Spec Ed	
Percent of students scoring competent or advanced on industry standards-based assessments (NOCTI and NIMS) and industry certifications earned.	Neg	Neg	Neg	Pos	Neg	Unknown
AP/IB/college course offerings in arts, English, history and social sciences, math and computer science, sciences, world languages and culture, and career pathways.	Neg	Neg	Neg	Pos	Neg	Neg
Postsecondary enrollment overall and within student subgroups or joining the military	Neg	Neg	Neg	Pos	Neg	Neg
Percent of graduates who enroll in college or join the military within 16 months of graduation.	Neg	Neg	Neg	Pos	Neg	Neg
Percent of graduates who are employed within 16 months of graduation.	Neg	Neg	Neg	Pos	Neg	Neg

"Neg" indicates that research strongly suggests this indicator would be negatively correlated with the specific school-level student characteristic."Pos" indicates that research strongly suggests this indicator would be positively correlated with the specific school-level student characteristic. Empty cells denote that there would likely not be a statistically significant correlation between the indicator and the school-level student characteristic.

f) Create Two SPP Scores that Address the Two Purposes

Given the low probability that any state could create an accountability system that adequately satisfies the aforementioned two purposes, we urge the state to report a score that for each of the two purposes. The first score would provide parents, educators, the public, and policymakers information about the levels of student achievement in each school in the Commonwealth. This is essentially the purpose that the SPP has served in the past and will likely serve in the future. The second score provide parents, educators, the public, and policymakers information about the effectiveness of schools in improving student performance apart from the influence of school-level student characteristics.

Even if not an official score, the public reporting of the second score serves multiple important purposes. First, the score would help parents identify the most effective schools in improving student outcomes, thus result in more informed choices for those families availing themselves of such options. Second, the information would assist educators and policymakers in identifying the most effective schools, thus allowing them to learn about the strategies and procedures employed by these schools. Third, and most importantly, the reporting of a school effectiveness score would reward schools and educators that often never receive any accolades but are certainly deserving of them.

g) Calculate More Accurate Achievement Gap Measures

The current methodology for identifying the achievement gap is not an accepted method for accurately identifying the achievement gap and provides inaccurate information about the achievement gap for a number of reasons. Any measure based on the percentage of students achieving proficiency will yield inaccurate and potentially misleading information about the achievement gap. PDE should explore various options for calculating the achievement gap and choose a method that accurately captures the true achievement gap. Using an inaccurate achievement gap will not only send incorrect signals to schools about their efforts to close the achievement gap, but using an inaccurate indicator will randomly confer unfair advantages and disadvantages to certain schools.

h) Adopt a Five-Year Graduation Rate or Adjust the Four-Year Graduation Rate

If allowed under ESSA, adopt a five year cohort graduation model. The current four year graduation cohort model penalizes: (1) schools that provide support for special education students continuing their education in a K-12 setting prior to entering the workforce or higher education; and, (2) schools that focus their efforts on recovering dropouts and helping them to obtain a high school diploma. In the first case, many students with specific disabilities such as autism may require additional years of schooling beyond graduation to be well-prepared for the workforce or for post-secondary schooling. In the latter case, many students that have dropped out of school may take longer than four years to obtain a diploma. The goal should be obtaining a diploma regardless of the time frame. In both cases, the Commonwealth should not punish schools that are trying to do what is right for their students. Other than completely moving to a five year graduation rate, PDE might adjust the calculation for students that have dropped out and for special education students.

i) Include a "Locally Selected Indicators" Section

Given the emphasis that PDE has placed on creating a "holistic" view of schooling, the Commonwealth should strongly consider allowing districts to choose from a menu of options. Such options could include, but not be limited to, the following:

- Percentage of students completing Advanced Placement, International Baccalaureate Diploma, or College Credit;
- Percentage of students participating in PSAT/Plan Participation;
- Percentage of students meeting SAT/ACT College Ready Benchmark;
- Percentage of students enrolling in post-secondary institutions of education within 18 months of graduation;
- Percentage of students finding full-time employment, including military service, within 18 months of graduation;
- Effective use of student engagement surveys;
- Effective use of school climate/culture surveys; and.
- Effective use of teacher working conditions surveys;

Unfortunately, most of these measures would be correlated with school-level student characteristics. However, this would be a choice that district leaders make. PDE could only provide adjusted results if data for every PA school were available for the indicator. We address the issue of surveys and how to appropriately incorporate them into the SPP is discussed below.

j) Encourage the Use of Survey Results

Given the recent evidence on the importance of non-cognitive outcomes such as student engagement[xxxv], student socio-emotional health, school climate[xxxvi], and teacher working conditions[xxxvii], PDE should strongly encourage schools to begin experimenting with use of such surveys as indicators of school quality. However, the amount of research on the use of such surveys for accountability purposes is currently insufficient for adoption for use in high-stakes systems such as the SPP. For example, we currently do not know if either teachers and/or students can devise strategies to game the system. More importantly, researchers are currently unclear as to whether we can accurately measure these outcomes in a manner accurate enough to include in school accountability systems.

Using survey results, however, must be done so with extreme caution. Low-response rates invalidate the accuracy of the results. Moreover, school personnel would certainly have the opportunity to manipulate the results by encouraging students to respond in particular ways. Thus, researchers suggest that the actual survey results not be used in accountability systems. Instead, researchers suggest that states collect and review a school's efforts at using the survey results to accurately identify issues and the plans to enact changes that address the issues. Schools, then, would have to submit detailed reviews of their use of the surveys and PDE would need to create working groups to review the plans and determine the number of points each school deserved. This would be a laborious and potentially expensive endeavor. However, the odds for a positive impact on schools and students is quite high, thus PDE should encourage this option.

Perhaps the best use of such surveys is in analyzing the strengths and weaknesses of schools as part of an in-depth screening of schools after schools are identified as being in the bottom 5% of schools using an intimal screening mechanism. Reviews teams could use such information to more accurately identify which schools are truly in need of assistance and which schools are already on a trajectory of more acceptable outcomes.

k) Replace the Use of Percent Proficient or Advanced with a Performance Index

Another drawback with the use of percent proficient or advanced as an accountability measures is that such a measure incentivizes educators to focus only on those students just below the cut point for proficient[xxxviii]. Indeed, schools are only rewarded for moving a student from not

proficient to proficient while not receiving any reward or recognition for moving a student from below basic to basic or from proficient to advanced. One strategy to address this concern is to create a performance index that rewards the advancement of students from one performance group to another. Importantly, the US Department of Education has approved the use of such an index under certain condition.

A Pennsylvania performance index could allot points to the percentage of students at each of four levels: below basic, basic, proficient, and advanced. To meet USDoE regulations, the points awarded must be greater for higher levels of performance. For example, as shown in Table 9, PDE could award 25 points for the percentage of students scoring Below Basic, 50 points for students scoring Basic, 75 points for scoring Proficient, and 100 points for scoring Advanced. Other points could be used as long as a school receives more points for students in the brackets for the higher levels of performance. For example, PDE could not award the same amount of points for students scoring Proficient and Advanced even if the percentages were the same. Thus, if a school had 30% of students scoring proficient and 30% of students scoring Advanced, the school must receive a greater number of points for the 30% of students scoring Advanced than for the 30% of students scoring proficient. Once the points are determined, PDE would then transform the points into accountability system points using a method similar to the methods used for other components of the accountability system.

Table 9: Example of Performance Index of Student PSSA and Keystone Performance

Achievement	School A			School B		
Level	%	Points	Award	%	Points	Award
Below Basic	20	25	500	35	25	875
Basic	40	50	2000	35	50	1750
Proficient	30	75	2250	25	75	1875
Advanced	10	100	1000	5	100	500
Total	100		5750	100		5000

III. CONCLUSION

As we have shown above, the prior School Performance Profile had some serious problems which created some unintended consequences for schools and the Commonwealth. Most importantly, the SPP clearly did not accurately identify school effectiveness if we define school effectiveness as *improving* the educational outcomes of students that attend a school. Instead, to a significant degree, the SPP identified the student demographics of schools and rewarded schools enrolling greater percentages of not economically disadvantaged students, not special education students, not ELL students, and not students of color. Because the SPP is a component in educator evaluations, the SPP helps create a incentives for educators to seek employment in schools that serve more affluent students, White students, students without disabilities, and students whose primary language is English. Thus, the SPP helps maintain the inequitable distribution of teachers across schools and districts that is an important factor in explaining the gap in achievement between economically disadvantaged and not economically disadvantaged students as well as between White students and students of color.

Unfortunately, the Future Ready PA proposal appears to have many of the same problems as the SPP. In particular, the Future Ready PA proposal includes a number of components that are highly correlated with student demographics. Thus, at this point in time, the Future Ready PA system would also not accurately identify school effectiveness and would continue to perpetuate the public school death spiral described above. We urge the Commonwealth to use the increased flexibility granted through ESSA to devise a system that is a far more accurate measure of school effectiveness than the SPP or the Future Ready PA system.

APPENDIX A: Correlation Between SPP Score/SPP Indicator Scores and School Characteristics for Lower Elementary Schools (AY 2013-14)

SPP Indicator	Statistics	Percentage of Students						School Size
		econ disadv	White or Asian	Female	ELL	Gifted	Spec Ed	
Calculated Score	Correlation	**-.752****	**.468****	-.049	-.090	.277**	-.177*	.084
	Stat Sig	**.000**	**.000**	.520	.238	.000	.019	.269
Final Academic Score	Correlation	**-.754****	**.467****	-.050	-.089	.282**	-.176*	.087
	Stat Sig	**.000**	**.000**	.508	.242	.000	.020	.255
Attendance Rate	Correlation	**-.635****	.227**	-.057	.071	.292**	-.257**	.027
	Stat Sig	**.000**	.003	.459	.353	.000	.001	.724
Promotion Rate	Correlation	-.149	-.038	.153*	.125	.178*	-.378**	.193*
	Stat Sig	.051	.623	.045	.102	.019	.000	.011
% Scoring Adv/Proficient on state tests: Math	Correlation	**-.672****	.350**	-.010	-.060	.266**	-.187*	.117
	Stat Sig	**.000**	.000	.900	.433	.000	.013	.124
% Scoring Adv/Proficient on state tests: Reading	Correlation	**-.742****	.468**	-.063	-.161*	.304**	-.131	.033
	Stat Sig	**.000**	.000	.411	.034	.000	.085	.667

Very Strong Correlation	Abs Value > 0.700
Strong Correlation	Abs Value: 0.500 to 0.699
Moderate Correlation	Abs Value: 0.300 to 0.499
Weak Correlation	Abs Value: < 0.300
No Correlation	Not Stat Significant

APPENDIX B Correlation Between SPP Score/SPP Indicator Scores and School Characteristics for Upper Elementary Schools (AY 2013-14)

SPP Indicator	Statistics	Percentage of Students						School Size
		econ disadv	White or Asian	Female	ELL	Gifted	Spec Ed	
Final Academic Score	Correlation	**-.769****	**.612****	.048	-.295**	.369**	-.125**	-.053*
	Stat Sig	**.000**	**.000**	.073	.000	.000	.000	.044
% Scoring Adv/Proficient on state tests: Math	Correlation	**-.831****	.726**	.029	-.351**	.374**	-.111**	-.080**
	Stat Sig	**0.000**	.000	.281	.000	.000	.000	.003
% Scoring Adv/Proficient on state tests: Reading	Correlation	**-.869****	.734**	.037	-.377**	.395**	-.121**	-.077**
	Stat Sig	**0.000**	.000	.161	.000	.000	.000	.003
% Scoring Adv/Proficient on state tests: Science	Correlation	**-.787****	.791**	.013	-.416**	.337**	-.026	-.115**
	Stat Sig	**.000**	.000	.628	.000	.000	.339	.000
% Scoring Adv/Proficient on state tests: Writing	Correlation	**-.699****	.494**	.088**	-.229**	.387**	-.142**	.025
	Stat Sig	**.000**	.000	.003	.000	.000	.000	.388
% Scoring Adv/Proficient on Grade 3 Reading	Correlation	**-.820****	.738**	.041	-.374**	.374**	-.075**	-.112**
	Stat Sig	**0.000**	.000	.129	.000	.000	.006	.000
Science Gap Closure All Students	Correlation	-.128**	.068*	.036	-.050	.065*	-.055*	-.018
	Stat Sig	.000	.012	.188	.065	.016	.043	.496
Science Gap Closure Hist Underperf Grps	Correlation	-.037	.033	.003	-.044	.002	-.014	-.023
	Stat Sig	.185	.239	.921	.119	.956	.628	.412
PVAAS Student Growth: Math	Correlation	-.103**	-.016	.013	.078**	.055*	-.088**	.039
	Stat Sig	.000	.535	.635	.003	.037	.001	.145
PVAAS Student Growth: Reading	Correlation	-.046	-.042	-.004	.066*	.060*	-.042	.052*
	Stat Sig	.082	.117	.871	.012	.023	.113	.048
PVAAS Student Growth: Science	Correlation	-.483**	.492**	.013	-.258**	.210**	.005	-.092**
	Stat Sig	.000	.000	.638	.000	.000	.850	.001
PVAAS Student Growth: Writing	Correlation	-.437**	.195**	.069*	-.085**	.276**	-.125**	.047
	Stat Sig	.000	.000	.018	.004	.000	.000	.107
Attendance Rate	Correlation	**-.685****	**.505****	.005	-.160**	.334**	-.130**	.000
	Stat Sig	**.000**	**.000**	.856	.000	.000	.000	.988
Promotion Rate	Correlation	-.368**	.251**	.026	.008	.194**	-.008	.013
	Stat Sig	.000	.000	.330	.758	.000	.775	.623
% Scoring Advanced on state tests: Math	Correlation	**-.839****	**.643****	.025	-.303**	.447**	-.129**	-.016
	Stat Sig	**0.000**	**.000**	.346	.000	.000	.000	.543
% Scoring Advanced on state tests: Reading	Correlation	**-.849****	**.615****	.030	-.300**	.488**	-.143**	-.005
	Stat Sig	**0.000**	**.000**	.259	.000	.000	.000	.850
% Scoring Advanced on state tests: Science	Correlation	**-.798****	**.693****	.008	-.351**	.381**	-.079**	-.092**
	Stat Sig	**.000**	**.000**	.767	.000	.000	.003	.001
% Scoring Advanced on state tests: Writing	Correlation	**-.546****	.299**	.066*	-.114**	.382**	-.170**	.055
	Stat Sig	**.000**	.000	.022	.000	.000	.000	.056

APPENDIX C Correlation Between SPP Score/SPP Indicator Scores and School Characteristics for Middle Schools (AY 2013-14)

SPP Indicator	Statistics	Percentage of Students						School Size
		econ disadv	White or Asian	Female	ELL	Gifted	Spec Ed	
Final_Academic_Score	Correlation	**-.801****	**.723****	.097*	-.378**	**.544****	-.243**	.117**
	Stat Sig	**.000**	**.000**	.012	.000	**.000**	.000	.003
% Scoring Adv/Proficient on state tests: Math	Correlation	**-.830****	**.797****	.114**	-.405**	**.536****	-.217**	.079*
	Stat Sig	**.000**	**.000**	.003	.000	**.000**	.000	.043
% Scoring Adv/Proficient on state tests: Reading	Correlation	**-.879****	**.826****	.115**	-.458**	**.578****	-.211**	.075
	Stat Sig	**.000**	**.000**	.003	.000	**.000**	.000	.054
% Scoring Adv/Proficient on state tests: Science	Correlation	**-.857****	**.793****	.081*	-.440**	.497**	-.258**	.099*
	Stat Sig	**.000**	**.000**	.036	.000	.000	.000	.011
% Scoring Adv/Proficient on state tests: Writing	Correlation	**-.798****	**.753****	.159**	-.413**	**.528****	-.211**	.059
	Stat Sig	**.000**	**.000**	.000	.000	**.000**	.000	.131
% Scoring Adv/Proficient on Grade 3 Reading	Correlation	**-.744****	**.696****	.189**	-.229**	.321**	-.144*	-.135
	Stat Sig	**.000**	**.000**	.009	.001	.000	.048	.063
Science Gap Closure All Students	Correlation	-.090*	.050	.024	-.036	.039	-.036	-.002
	Stat Sig	.023	.208	.551	.364	.329	.365	.965
Science Gap Closure Hist Underperf Grps	Correlation	-.074	.050	-.054	-.030	.068	.009	-.004
	Stat Sig	.062	.206	.174	.447	.087	.829	.923
PVAAS Student Growth: Math	Correlation	-.068	-.024	.004	.098*	.148**	-.165**	.150**
	Stat Sig	.082	.534	.925	.011	.000	.000	.000
PVAAS Student Growth: Reading	Correlation	-.049	-.033	.037	.044	.148**	-.159**	.120**
	Stat Sig	.210	.391	.344	.258	.000	.000	.002
PVAAS Student Growth: Science	Correlation	**-.651****	**.634****	-.021	-.348**	.389**	-.095*	.072
	Stat Sig	**.000**	**.000**	.589	.000	.000	.014	.063
PVAAS Student Growth: Writing	Correlation	-.457**	.366**	.152**	-.207**	.321**	-.177**	.060
	Stat Sig	.000	.000	.000	.000	.000	.000	.121
Attendance Rate	Correlation	**-.641****	**.521****	.081*	-.225**	.373**	-.206**	.041
	Stat Sig	**.000**	**.000**	.039	.000	.000	.000	.296
Promotion Rate	Correlation	-.383**	.329**	.012	-.070	.272**	-.002	.085*
	Stat Sig	.000	.000	.762	.075	.000	.960	.031
% Scoring Advanced on state tests: Math	Correlation	**-.853****	**.777****	.073	-.373**	**.582****	-.223**	.132**
	Stat Sig	**.000**	**.000**	.062	.000	**.000**	.000	.001
% Scoring Advanced on state tests: Reading	Correlation	**-.893****	**.797****	.069	**-.423****	**.639****	-.210**	.120**
	Stat Sig	**.000**	**.000**	.077	**.000**	**.000**	.000	.002
% Scoring Advanced on state tests: Science	Correlation	**-.822****	**.716****	.021	-.381**	**.485****	-.249**	.098*
	Stat Sig	**.000**	**.000**	.588	.000	**.000**	.000	.012
% Scoring Advanced on state tests: Writing	Correlation	**-.695****	**.553****	.069	-.270**	**.508****	-.189**	.145**
	Stat Sig	**.000**	**.000**	.075	.000	**.000**	.000	.000

APPENDIX D Correlation Between SPP Score/Primary SPP Indicator Scores and School Characteristics for High Schools (AY 2013-14)

SPP Indicator	Statistics	Percentage of Students						School Size
		econ disadv	White or Asian	Female	ELL	Gifted	Spec Ed	
Final Academic Score	Correlation	**-.780****	**.618****	.155**	-.396**	.368**	-.409**	.160**
	Stat Sig	**.000**	**.000**	.000	.000	.000	.000	.000
% Scoring Adv/Proficient on state tests: Math	Correlation	**-.732****	**.663****	.197**	-.461**	.387**	-.447**	.069
	Stat Sig	**.000**	**.000**	.000	.000	.000	.000	.075
% Scoring Adv/Proficient on state tests: Reading	Correlation	**-.741****	**.647****	.254**	-.468**	.410**	-.454**	.088*
	Stat Sig	**.000**	**.000**	.000	.000	.000	.000	.023
% Scoring Adv/Proficient on state tests: Science	Correlation	**-.790****	**.717****	.112**	-.455**	.338**	-.355**	.106**
	Stat Sig	**.000**	**.000**	.004	.000	.000	.000	.006
% Scoring Adv/Proficient on state tests: Writing	Correlation	**-.546****	.496**	.057	-.395**	.318**	-.458**	-.232**
	Stat Sig	**.000**	.000	.421	.000	.000	.000	.001
% Scoring Adv/Proficient on Grade 3 Reading	Correlation	**-.549****	**.547****	.012	-.180	.093	.032	-.092
	Stat Sig	**.003**	**.003**	.951	.369	.643	.874	.647
Science Gap Closure All Students	Correlation	-.277**	.235**	.072	-.183**	.132**	-.163**	.051
	Stat Sig	.000	.000	.063	.000	.001	.000	.187
Science Gap Closure Hist Underperf Grps	Correlation	-.190**	.178**	.069	-.157**	.111**	-.104**	-.004
	Stat Sig	.000	.000	.079	.000	.005	.008	.913
PVAAS Student Growth: Math	Correlation	-.398**	.249**	.064	-.189**	.094*	-.169**	.078*
	Stat Sig	.000	.000	.101	.000	.016	.000	.043
PVAAS Student Growth: Reading	Correlation	-.373**	.191**	.144**	-.181**	.144**	-.215**	.113**
	Stat Sig	.000	.000	.000	.000	.000	.000	.003
PVAAS Student Growth: Science	Correlation	**-.516****	.386**	.024	-.224**	.179**	-.158**	.154**
	Stat Sig	**.000**	.000	.529	.000	.000	.000	.000
PVAAS Student Growth: Writing	Correlation	-.097	.009	.107	-.127	.144*	-.227**	-.161*
	Stat Sig	.172	.898	.132	.073	.041	.001	.023
Attendance Rate	Correlation	**-.578****	**.500****	.117**	-.365**	.231**	-.411**	.079*
	Stat Sig	**.000**	**.000**	.002	.000	.000	.000	.040
Cohort Graduation Rate	Correlation	**-.532****	.435**	.099*	-.410**	.263**	-.386**	-.021
	Stat Sig	**.000**	.000	.011	.000	.000	.000	.591

APPENDIX E Correlation Between SPP Score/Primary SPP Indicator Scores and School Characteristics for High Schools (AY 2013-14)

SPP Indicator	Statistics	Percentage of Students						School Size
		econ disadv	White or Asian	Female	ELL	Gifted	Spec Ed	
College Ready Benchmark (SAT or ACT)	Correlation	**-.810****	.590**	.137**	-.342**	.420**	-.414**	.175**
	Stat Sig	**.000**	.000	.000	.000	.000	.000	.000
% Scoring Adv/Proficient on IBSCA	Correlation	-.315**	.263**	.154**	-.185**	.050	-.125*	.021
	Stat Sig	.000	.000	.003	.000	.331	.014	.688
% Scoring Advanced on state tests: Math	Correlation	-.445**	.413**	.134**	-.288**	.333**	-.273**	.029
	Stat Sig	.000	.000	.000	.000	.000	.000	.452
% Scoring Advanced on state tests: Reading	Correlation	-.384**	.393**	.133**	-.273**	.310**	-.258**	.000
	Stat Sig	.000	.000	.001	.000	.000	.000	.992
% Scoring Advanced on state tests: Science	Correlation	**-.626****	.465**	.063	-.272**	.381**	-.257**	.174**
	Stat Sig	**.000**	.000	.100	.000	.000	.000	.000
% Scoring Advanced on state tests: Writing	Correlation	-.423**	.253**	.136	-.198**	.378**	-.344**	-.071
	Stat Sig	.000	.000	.054	.005	.000	.000	.315
% Scoring Advanced on ISBCA	Correlation	-.363**	.360**	.111*	-.255**	.089	-.105*	-.028
	Stat Sig	.000	.000	.030	.000	.082	.041	.588
% Scoring >=3 om AP/>4 on IB	Correlation	**-.658****	.367**	.124**	-.157**	.437**	-.284**	.321**
	Stat Sig	**.000**	.000	.003	.000	.000	.000	.000

APPENDIX F:
Relationship between the Percent of Economically Disadvantaged Students and SPP Scores by School Level

Figure F1: Elementary Schools (2014)

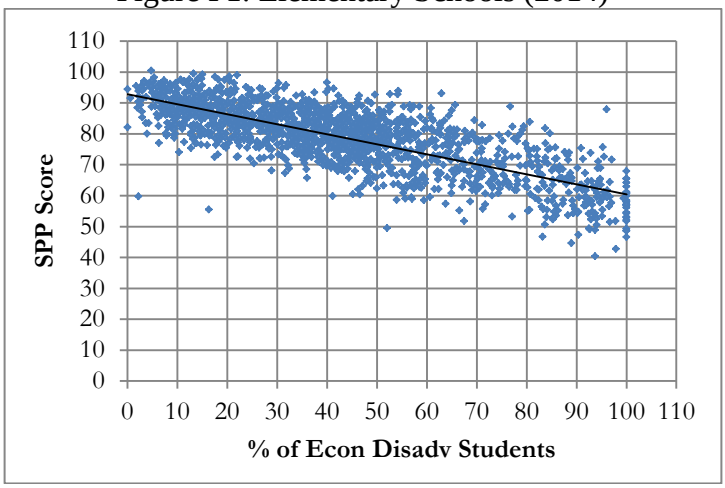


Figure F2: Middle Schools (2014)

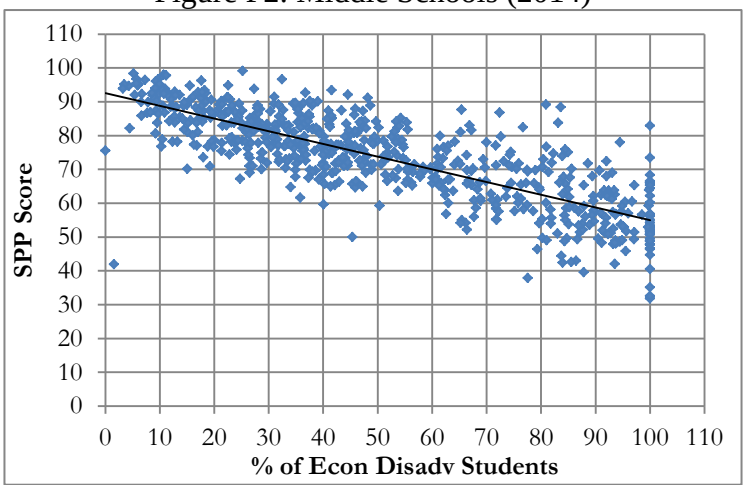
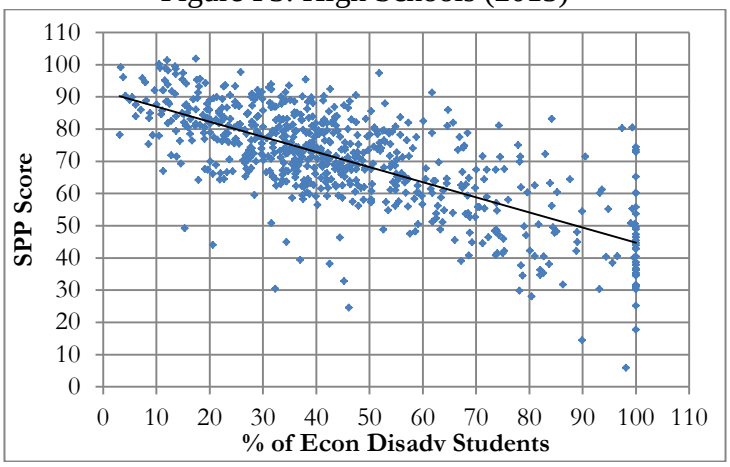


Figure F3: High Schools (2015)



About CEEPA and Dr. Ed Fuller

The Penn State Center for Evaluation and Policy Analysis (CEEPA) is the College of Education's most recent additions to its portfolio of research and evaluation centers. The Center's mission is to provide unbiased, high-quality evaluation and policy analysis services to education and other organizations in the Commonwealth of Pennsylvania and across the nation. The Director of the Center is Dr. Ed Fuller. Dr. Fuller is an associate professor in the Education Policy Studies Department in Penn State's College of Education. He is also the Executive Director of the Center for Evaluation and Education Policy Analysis at Penn State. He has three degrees in education, including a master's degree in school administration and PhD in Education Policy, both from the University of Texas at Austin. Dr. Fuller has 25 years of experience as an educator in a variety of roles, including teacher, Director of Research and Evaluation at the Texas State Board for Educator Certification, program evaluator, consultant for the Texas legislature, and professor. Dr. Fuller may be contacted by email at ejf20@psu.edu, by phone at 814-865-2233, or by cell phone at 512-971-5715

---

[i] Braun, H. (2004). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives, 12*(1), 1-43; Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*(24), 305-331.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*(9-10), 1045-1057.

Dee, T., & Jacob, B. (2011). The Impact of No Child Left Behind on Student Achievement. *Journal of Policy Analysis and Management, 30*(3), 418-446.

Figlio, D., & Rouse, C. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics, 90*(1-2), 239-255.

Hanushek, E., & Raymond, M. (2005). Does school accountability lead to imprvoed student performance? *Journal of Policy Analysis and Management*(24), 297-327.

Reback, R., Rockoff, J., & Schwartz, H. (2014). Under pressure: Job security, resource allocation, and productivity in schools under NCLB. *American Economic Journal: Economic Policy, 6*(3).

Winters, M., Trivitt, J., & Greene, J. (2010). The impact of high-stakes testing on student proficiency in low-stakes subjects: Evidence from Florida's elementary science exam. *Economics of Education Review, 29*(1), 138-146.

[ii] Davidson, E., Reback, R., Rockoff, J. E., & Schwartz, H. L. (2015). Fifty ways to leave a child behind: Idiosyncrasies and discrepancies in states' implementation of NCLB. Educational Researcher, 44(6), 347–358; Polikoff, McEachin, Wrabel, & Duque, 2013.

[iii] Data Recognition Corporation. (2005). Technical Report for the Pennsylvania System of School Assessment: 2005 Reading and Mathematics. p.2 Retrieved from: http://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PSSA/Technical%20Reports/2005%20PSSA%20Reading%20and%20Math%20Technical%20Report.pdf

[iv] Jennings, J., & Sohn, H. (2014, p. 138). Measure for measure: How proficiency-based accountability systems affect inequality in academic achievement. Sociology of Education, 87(2), 125–141.

Polikoff, M., McEachin, A. J., Wrabel, S. L., & Duque, M. (2013, p.1). The wave of the future? School accountability in the waiver era. Educational Researcher, 43(1), 45–54.

[v] Braun, H. (2004). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives, 12*(1), 1-43; Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*(24), 305-331.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*(9-10), 1045-1057.

Dee, T., & Jacob, B. (2011). The Impact of No Child Left Behind on Student Achievement. *Journal of Policy Analysis and Management, 30*(3), 418-446.

Figlio, D., & Rouse, C. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics, 90*(1-2), 239-255.

Hanushek, E., & Raymond, M. (2005). Does school accountability lead to imprvoed student performance? *Journal of Policy Analysis and Management*(24), 297-327.

Reback, R., Rockoff, J., & Schwartz, H. (2014). Under pressure: Job security, resource allocation, and productivity in schools under NCLB. *American Economic Journal: Economic Policy, 6*(3).

Winters, M., Trivitt, J., & Greene, J. (2010). The impact of high-stakes testing on student proficiency in low-stakes subjects: Evidence from Florida's elementary science exam. *Economics of Education Review, 29*(1), 138-146.

[vi] Davidson, E., Reback, R., Rockoff, J. E., & Schwartz, H. L. (2015). Fifty ways to leave a child behind: Idiosyncrasies and discrepancies in states' implementation of NCLB. Educational Researcher, 44(6), 347–358; Polikoff, McEachin, Wrabel, & Duque, 2013.

[vii] Fitzpatrick, J., Sanders, J. R., & Worthen, B. R. (2011). *Program Evaluation: Alternative Approaches and Practical Guidelines* (4th ed.). New Jersey: Pearson Education.

[viii] Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

[ix] Grissom, J., Kalgorides, D., & Loeb, S. (2012, November). "Using student test scores to measure principal performance." Working paper No. 18568. Cambridge, MA: National Bureau of Economic Research.

Kane, T.J. & Staiger, D.O. (2002). "The promise and pitfalls of using imprecise school accountability measures," *The Journal of Economic Perspectives,* 16, 91-114.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

[x] Fitzpatrick, J., Sanders, J. R., & Worthen, B. R. (2011). *Program Evaluation: Alternative Approaches and Practical Guidelines* (4th ed.). New Jersey: Pearson Education.

Koretz, D. M. (2008). *Measuring up*. Harvard University Press.

[xi] Erdogan, B. (2002). "Antecedents and consequences of justice perceptions in performance appraisals." *Human Resource Managemewnt Review, 12*: 555-578.

Kane, T.J. & Staiger, D.O. (2002). "The promise and pitfalls of using imprecise school accountability measures," *The Journal of Economic Perspectives,* 16, 91-114.

[xii] Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American educational research journal*, *42*(2), 231-268.

[xiii] Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American educational research journal*, *42*(2), 231-268.

[xiv] Koretz, D. M. (2008). *Measuring up*. Harvard University Press.

[xv] Michelmore, K., & Dynarski, S. (2016). *The Gap within the Gap: Using Longitudinal Data to Understand Income Differences in Student Achievement* (No. w22474). National Bureau of Economic Research.

[xvi] Ladd, H. F. (2012). Education and poverty: Confronting the evidence. *Journal of Policy Analysis and Management*, *31*(2), 203-227.

[xvii] Ladd, H. F. (2012). Education and poverty: Confronting the evidence. *Journal of Policy Analysis and Management*, *31*(2), 203-227.

Morsy, L., & Rothstein, R. (2015). Five Social Disadvantages That Depress Student Performance: Why Schools Alone Can't Close Achievement Gaps. Report. *Economic Policy Institute*.

Slates, S. L., Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2012). Counteracting summer slide: Social capital resources within socioeconomically disadvantaged families. *Journal of Education for Students Placed at Risk (JESPAR)*, *17*(3), 165-185.

[xviii] Ladd, H. F. (2012). Education and poverty: Confronting the evidence. *Journal of Policy Analysis and Management*, *31*(2), 203-227.

Morsy, L., & Rothstein, R. (2015). Five Social Disadvantages That Depress Student Performance: Why Schools Alone Can't Close Achievement Gaps. Report. *Economic Policy Institute*.

Yoshikawa, H., Aber, J. L., & Beardslee, W. R. (2012). The effects of poverty on the mental, emotional, and behavioral health of children and youth: implications for prevention. *American Psychologist*, *67*(4), 272.

[xix] http://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PVAAS/Professional%20Development/PVAAS%20Misconceptions%20Booklet.pdf

[xx] Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. J. (2014). The sensitivity of value-added estimates to specification adjustments: Evidence from school-and teacher-level models in Missouri. *Statistics and Public Policy*, *1*(1), 19-27.

[xxi] Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. J. (2014). The sensitivity of value-added estimates to specification adjustments: Evidence from school-and teacher-level models in Missouri. *Statistics and Public Policy*, *1*(1), 19-27. Page 26.

[xxii] Harris, D., & Anderson, A. (2013). Does value-added work better in elementary than in secondary grades?. *Washington, DC: Carnegie Knowledge Netowrk*

[xxiii] Holloway-Libell, J. (2015). Evidence of grade and subject-level bias in value-added measures. *Teachers College Record*. http://www.tcrecord.org/Content.asp?ContentID=17987

[xxiv] Harris, D. N., & Herrington, C. D. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American journal of education*, *112*(2), 209-238.

[xxv] Harris, D. N., & Herrington, C. D. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American journal of education*, *112*(2), 209-238.

[xxvi] DiCarlo, M. (2014, November 17). Rethinking the use of simple achievement gap measures in school accountability systems. Shanker Blog, Albert Shanker Institute. Retrieved at: http://www.shankerinstitute.org/blog/rethinking-use-simple-achievement-gap-measures-school-accountability-systems

Ho, A. D., & Reardon, S. F. (2012). Estimating achievement gaps from test scores reported in ordinal "proficiency" categories. *Journal of Educational and Behavioral Statistics*, *37*(4), 489-517.

[xxvii] Ladd, H. F. (2001). School—Based Educational Accountability Systems: The Promise and the Pitfalls. *National Tax Journal*, *54*(2), 385-400.

Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: getting the incentives right. *Economics of Education review*, *21*(1), 1-17.

[xxviii] Figlio, D., & Loeb, S. (2011). School accountability. In Hanushek, E. A., Machin, S. J., & Woessmann, L. (Eds.). *Handbook of the Economics of Education*, *3*(8), 383-417. Elsevier. St. Louis, MO.

Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: getting the incentives right. *Economics of Education review*, *21*(1), 1-17.

[xxix] Figlio, D., & Loeb, S. (2011). School accountability. In Hanushek, E. A., Machin, S. J., & Woessmann, L. (Eds.). *Handbook of the Economics of Education*, *3*(8), 383-417. Elsevier. St. Louis, MO.

[xxx] Figlio, D., & Loeb, S. (2011). School accountability. In Hanushek, E. A., Machin, S. J., & Woessmann, L. (Eds.). *Handbook of the Economics of Education*, *3*(8), 383-417. Elsevier. St. Louis, MO.

[xxxi] Ladd, H. F. (2001). School—Based Educational Accountability Systems: The Promise and the Pitfalls. *National Tax Journal*, *54*(2), 385-400.

[xxxii] Feng, L., Figlio, D. N., & Sass, T. (2010). *School accountability and teacher mobility* (No. w16070). National Bureau of Economic Research.

Figlio, D., & Loeb, S. (2011). School accountability. In Hanushek, E. A., Machin, S. J., & Woessmann, L. (Eds.). *Handbook of the Economics of Education*, *3*(8), 383-417. Elsevier. St. Louis, MO.

[xxxiii] Figlio, D., & Loeb, S. (2011). School accountability. In Hanushek, E. A., Machin, S. J., & Woessmann, L. (Eds.). *Handbook of the Economics of Education*, *3*(8), 383-417. Elsevier. St. Louis, MO.

[xxxiv] Pennsylvania School Boards Association (2016, Jan.). *State Board of Education Receives Update on Proposed SPP Changes*. Retrieved from https://www.psba.org/2016/01/state-board-spp-changes/

[xxxv] Christenson, S. L., Reschly, A. L., & Wylie, C. (Eds.). (2012). *Handbook of research on student engagement*. Springer Science & Business Media.

[xxxvi] Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research*, *83*(3), 357-385.

Wang, M. T., & Degol, J. L. (2015). School climate: A review of the construct, measurement, and impact on student outcomes. *Educational Psychology Review*, 1-38.

[xxxvii] Kraft, M., Papay, J. P., Charner-Laird, M., Johnson, S. M., Ng, M., & Reinhorn, S. K. (2013). *Committed to their students but in need of support: How school context influences teacher turnover in high-poverty*. urban schools. Working Paper: Project on the Next Generation of Teachers. Cambridge, MA: Harvard Graduate School of Education.

Ladd, H. F. (2011). Teachers' Perceptions of Their Working Conditions How Predictive of Planned and Actual Teacher Movement?. *Educational Evaluation and Policy Analysis*, *33*(2), 235-261.

[xxxviii] Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American educational research journal*, *42*(2), 231-268.