AMERICAN INSTITUTES FOR RESEARCH®

The Every Student Succeeds Act in Pennsylvania:

Recommendations From Stakeholder Work Groups and Associated Research

OCTOBER 2016

IMPORTANT NOTE FROM THE PENNSYLVANIA DEPARTMENT OF EDUCATION: The workgroup recommendations described in this report reflect the work and consensus of each stakeholder work group. These recommendations will be carefully evaluated by the Department, and considered in the context of relevant evidence, research, and additional stakeholder engagement as Pennsylvania develops its ESSA State Plan.

The Every Student Succeeds Act in Pennsylvania:

Recommendations From Stakeholder Work Groups and Associated Research

October 2016

This report was prepared for the Pennsylvania Department of Education by multiple authors from the American Institutes for Research. Mariann Lemke and Katelyn Lee authored the assessment section; Kerstin LeFloch and David English authored the accountability section; Lynn Holdheide and Jeremy Rasmussen authored the educator preparation section; and Ellen Sherratt and Cassandra Meyer authored the educator evaluation section.

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000

www.air.org

7333_010/16

Acknowledgements

The Pennsylvania Department of Education wishes to acknowledge the contributions of the following individuals and organizations whose efforts were essential to the developing the recommendations found in this report and to the report itself:

ESSA Work Groups

Eighty-two individuals, identified via nominations from more than 60 stakeholder organizations, contributed their time and expertise as members of the four work groups charged with creating the recommendations described in this report. A full list of each work group's membership can be found in Appendix A or online at the Pennsylvania Department of Education's ESSA website at http://www.education.pa.gov/Pages/Every-Student-Succeeds-Act.aspx.

Council of Chief State School Officers

Staff and advisors from the Council of Chief State School Officers (CCSSO) were essential in facilitating work group conversations. In particular, Scott Norton facilitated the assessment work group, Alissa Peltzman the accountability work group, David Hendrie the educator preparation work group, and Jeanne Harmon the educator evaluation work group. The Department also thanks Peter Zamora, Director of Federal Relations at CCSSO, for his leadership.

The Consortium for Policy Research in Education

Researchers from the Consortium for Policy Research in Education (CPRE)—including Richard Ingersoll, Bobbi Newman, Matthew Steinberg, and Jonathan Supovitz—provided valuable policy context and research to Department staff.

Contents

Page

Acknowledgements i

Executive Summary iii

Introduction 1

Assessment 3

ESSA Requirements 3

Overview of Assessment Work Group Recommendations 3

Accountability 13

Summary of measures by type 24

Educator Preparation 41

ESSA Requirements 41

Educator Evaluation 61

Summary of Findings 61

ESSA Requirements 61

References 70

Appendix A: List of Stakeholders A-1

Appendix B: Links to Additional Resources B-1

Executive Summary

On December 10, 2015, President Obama signed the Every Student Succeeds Act (ESSA), which reauthorized the Elementary and Secondary Education Act (ESEA). Developed and passed with strong, bipartisan agreement, ESSA replaces the No Child Left Behind Act (NCLB) and provides states and communities with new flexibility to manage federal education policy. ESSA requires that states develop and submit a State Plan to the U.S. Department of Education; states have the option of expanding these plans to address other important areas of federal education policy.[1]

The Pennsylvania Department of Education (PDE) believes that educator and stakeholder voice is crucial to the development of a coherent, and ultimately successful, State Plan. To ensure a solid foundation for State Plan development and further stakeholder engagement, the Department convened four work groups that explored the following components of the new law:

- Assessment
- Accountability
- Educator preparation[2]
- Educator evaluation

Eighty-two individuals, identified via nominations from more than 60 stakeholder organizations, contributed their time and expertise as members of these work groups. Each work group was charged with developing three to five recommendations to be considered by the Department and other education policymakers as appropriate in the development and implementation of the State Plan.

The Department contracted with the American Institutes for Research (AIR) to independently summarize the work groups' recommendations and relevant state policy and research. This approach aims to ensure that the Department and others account for the experiences of other states and systems that have implemented similar reforms and practices. Further, ESSA places significant emphasis on evidence-based practices, and this report is a first step in grounding Pennsylvania's State Plan in rigorous research and relevant policy analysis.

In some cases, available research did not match the specifics of the recommendation or addressed only part of the recommendation. In many other cases, it is difficult to generalize from the research to the recommendation because the recommendation is very broad. That is, the specifics of how recommendations are implemented will clearly have a strong bearing on the extent to which they can achieve successful outcomes. In these cases, we present information that is relevant and highlight how it relates to the recommendation. The Department will continue to work with all interested stakeholders to develop a State Plan that will best serve Pennsylvania's students, educators, and communities and reflects the best available research.

[1] States may submit plans in March or July 2017.

[2] Note that the educator preparation work group was originally charged with considering educator certification; however, because the group's discussion broadened beyond certification requirements alone, in this report we refer it as "educator preparation."

Assessment

Recommendation 1. PDE should reduce ESSA-required, statewide testing time for all students.

Overall, there is little evidence to suggest that reducing time spent on statewide assessments will lead to improved outcomes for students. Although students can benefit from increased instructional time, research on these benefits generally focuses on significant increases in instructional time from extending the school year, school hours, or afterschool time. It is not clear that any time saved by decreasing required testing would be used for instruction nor that the amount of time saved (likely a few hours) would be sufficient to have any effects.

However, it is likely that districts are spending time not reflected in state test administration time on activities related to state testing, such as test preparation. Reducing these activities and decreasing the emphasis on the use of test results within accountability policies have the potential to influence instruction and student outcomes.

Recommendation 2. Pursuant to decreasing the time spent on ESSA-required, statewide assessments, PDE should conduct a study to determine the feasibility of administering assessments at multiple points in time to better inform instruction.

To the extent that PDE wishes to investigate the possibilities of administering assessments at multiple points in time to better inform instruction, current research seems to suggest that though such an approach is technically feasible and perhaps desirable from the standpoint of providing more frequent information, it may necessitate significant investments in professional learning along with investments in development of such a system itself.

Recommendation 3. PDE should utilize a standards-aligned, state-required multiple choice-only assessment to meet ESSA requirements. PDE should encourage local education agencies (LEAs) to utilize performance-based measures for students to demonstrate progress toward achievement of postsecondary goals.

Evidence accumulated over the last 15 to 20 years suggests that although using a multiple-choice-only assessment may offer financial benefits, it may also carry risks in terms of potentially influencing undesirable instructional or other practices in schools and districts. Performance-based assessment, though holding promise in terms of its relationship to instruction, may require significant investments to ensure that it can be implemented successfully.

Accountability

Recommendation 1. The accountability system should start with a student-centered approach which considers the whole student experience including academics, physical and cultural environment and supports.

Research supports the link between inputs that support the whole child and academic success, but there is little evidence of their efficacy in the context of identifying schools for accountability purposes.

Recommendation 2. The PA accountability system should be based on an array of indicators of student experiences and growth toward college and career readiness, appropriately selected and weighted to serve different purposes, including:

- ***Identifying schools for ESSA supports, intervention, and recognition;***
- ***Timely reporting of meaningful information to schools, policymakers, and communities; and***
- ***Setting statewide, school, and community goals and interim targets.***

Using multiple measures can increase the validity and reliability of overall accountability determinations and support a richer theory of action for identifying leverage points for school improvement. However, despite consensus among policymakers and researchers about the importance of using multiple measures in accountability systems, there is little research to support decisions regarding which exact measures to use or how best to combine them. Some states are already including indicators related to college and career readiness in their accountability systems, which could serve as examples for Pennsylvania to consider. In addition, there are a number of examples of different state approaches to indicators that are required under ESSA, such as achievement status and growth, which may also be useful for Pennsylvania to consider.

Researchers do note that particular measures may be more suited for one role or another based on their technical validity, transparency, or other characteristics. Policymakers should consider the trade-offs between transparency, accuracy, fairness and potential for corruption as they consider indicators under ESSA. Fairness, in particular, must be considered from the perspective of each of the relevant stakeholders, such as students, parents, schools, and educators.

Recommendation 3. The PA accountability system will enable system wide continuous and sustainable improvement by providing transparent, timely, and meaningful feedback to all stakeholders.

To enable continuous improvement, careful consideration of how accountability results are reported is critical. To provide a strong and clear message that motivates stakeholders, a single summative score or rating might be best; a dashboard type of approach, however, may provide more insight into strengths and weaknesses and more readily support school improvement. There is some evidence that different stakeholders, such as parents and schools, prefer different approaches. Here, too, a number of existing state practices might provide examples for Pennsylvania to consider.

To support improvement efforts, however, research clearly suggests that states must ensure an adequate data infrastructure, be timely in delivering results, provide time for educators to use data, and build their capacity to do so.

Recommendation 4. The interventions in Pennsylvania's accountability system are evidence-based and applied in ways that are flexible and responsive to varying needs of students, communities, and schools to support the growth of every child. Pennsylvania's system includes a framework for district differentiated recognition, accountability and assistance. The level of state response is dependent on the tier status of the LEA. The tiered system classifies schools and LEAs on multiple levels based on multiple measures. The level or tier indicates the amount and type of support/intervention needed to improve student outcomes.

There are insufficient causal studies to provide a clear roadmap for states seeking to redesign their system of supports. However, over the past decade, scholars and practitioners have attempted to synthesize lessons learned from research and practice. Some of these tenets, such as providing significant resources to support planning and treating the district as the unit of change, are directly applicable to Pennsylvania's current work.

Educator Preparation

Recommendation 1. The Department should promote and increase opportunities to recruit, retain, and ensure a diverse, talented, and supported educator workforce.

The work group recommended a series of sub-recommendations related to this more general recommendation, as follows:

- **Sub-Recommendation 1a:** Promote and market teaching as a valued and respected profession;
- **Sub-Recommendation 1b:** Improve recruitment efforts through the use of financial incentives and by targeting diverse populations;
- **Sub-Recommendation 1c:** Investigate certification requirements considering quality and effect on diversity recruitment; and,
- **Sub-Recommendation 1d:** Strengthen educator support across the career continuum.

At this point little research exists on the success of efforts to promote and market teaching, though there are examples of programs to improve perceptions and increase recruitment into the profession which may serve as useful examples for Pennsylvania to consider.

Research about teacher compensation continues to suggest that salaries affect the labor market decisions that teachers make. Findings related to the use of incentives suggest a mixed level of success in teacher recruitment and retention. There is also research that a diverse teaching force may improve student achievement. Current research suggests that basic skills tests have disproportionate effects on minority candidates, potentially creating a barrier to minority populations pursuing teaching as a profession.

Research supports that induction and mentoring can have positive effects on teacher retention and improvements in practice; however, success is largely dependent on the quality of the induction and mentoring programs. While using educator evaluation data to guide professional learning and support provided through induction and mentoring is logical, research supporting this recommendation is not yet available.

Recommendation 2. The Department will define effective teachers as those who strive to engage all students in learning, demonstrate instructional and subject matter competence, and continuously grow and improve.

Though significant research has been done on measuring effective teaching, definitions of effective teaching or an effective educator center on theory and beliefs about what makes for successful teaching. Pennsylvania's 2015 Equity Plan simply defines "effective" educators as those whose overall effectiveness rating is "proficient" or "distinguished."

Many states and professional organizations have created their own definitions, which generally speak to multiple elements, such as teachers' contributions to student learning and other student outcomes, their contributions to their profession, knowledge of and practice of teaching, and possibly also relationship with parents or the community.

Recommendation 3. The Department should promote and support collaborative in-field, practical experiences as a crucial component of educator preparation.

Some research and policies support the idea of strong partnerships between IHEs and districts to improve teacher candidate quality. Research also shows that having a strong mentor or cooperating teacher can positively impact a teacher candidate, though specific strategies how to improve the training, expectations and incentives for cooperating teachers is still emerging. Finally, there is research on the importance of quality clinical training experiences, but there is less research on exactly what those programs should look like.

Recommendation 4. The Department should promote and increase opportunities to recruit, retain, and support diverse and talented school leaders.

There are some existing policy recommendations on principal recruitment and examples of programs which may serve as examples for Pennsylvania to consider, but there is little direct evidence on the best strategies or practices to promote and retain principals.

Although significant research may support the claim that effective principals are critical, there is limited evidence about how best to support principals with coaching or mentoring. There is some research which suggests that a core set of principal leadership practices, ranging from human capital management to agenda setting to coaching and instructional leadership, are associated with improved student outcomes, but also research which indicates that few principals actually engage in these practices. There is some limited research indicating that intensity of professional development may be important to bring about meaningful changes in principal effectiveness.

Educator Evaluation

Recommendation 1: Revise the overall components of the professional evaluation systems to reflect the following provisions that support teacher quality and student achievement: 80% professional practice (observation) and 20% student measures (SPP or combination of SPP and other relevant data as identified in the LEA's comprehensive plan).

The Educator Evaluation work group's recommendations to include only two measures and weight the professional practice measures at 80% of an educator's rating may run counter to the best available research. However, given limitations of the research base and the importance of stakeholder support, such changes might best align the educator evaluation system with the values of educators and other stakeholders.

Recommendation 2. Ensure that LEAs implement PA's educator evaluation system using a differentiated and collaborative process which promotes educator growth.

The work group recommended a series of sub-recommendations related to this more general recommendation, as follows:

- **Sub-Recommendation 2a:** Include position-specific observation rubrics in the educator evaluation system
- **Sub-Recommendation 2b:** Rotate educators with no performance concerns through cycles of formal evaluation and supportive growth
- **Sub-Recommendation 2c:** Assure evaluator competence in the use of observation rubrics
- **Sub-Recommendation 2d:** Provide timely, formative feedback

With respect to the idea of position-specific rubrics and rotating educators through cycles of evaluation, there is no extant research base related to these specific practices in education, but there are examples from other states. These examples may serve as useful guidance for Pennsylvania, and such practices could be relevant for stakeholder support of the system.

In contrast, research supports the notion that evaluator competence is important, and offers information on specific practices, including initial training, certification, use of multiple observers and conducting system reliability checks.

The importance of ensuring that evaluations result in timely feedback for teachers is supported by the existing research base. Indeed, research suggests some evidence about specifics of feedback: the value of keeping feedback focused on the task, not the learner (or self); employing a rubric that can clearly demonstrate the alignment between the teacher's actions and the desired goal (reduce uncertainty between performance and goals); focusing a few high leverage behaviors so that feedback can be delivered in manageable units; aligning with the district's and school's vision of teaching so that, overall, the teacher does not get conflicting feedback; allowing opportunities for practice between sessions so that feedback can be delivered after the

teacher has attempted a solution; and establishing a committed relationship between teacher and coach so that the teacher is more open to processing negative feedback. Translating this evidence very specifically to the Pennsylvania context and successfully implementing such a feedback system, however, may continue to present challenges.

Introduction

On December 10, 2015, President Obama signed the Every Student Succeeds Act (ESSA), which reauthorized the Elementary and Secondary Education Act (ESEA). Developed and passed with strong, bipartisan agreement, ESSA replaces the No Child Left Behind Act (NCLB) and provides states and communities with new flexibility to manage federal education policy. ESSA requires that states develop and submit a State Plan to the U.S. Department of Education; states have the option of expanding these plans to address other important areas of federal education policy.[3]

To ensure that Pennsylvania's State Plan is rooted in the day-to-day needs of educators, students, and communities, the Pennsylvania Department of Education (PDE) designed a stakeholder engagement process that began with four work groups that explored the following components of the new law:

- Assessment
- Accountability
- Educator preparation[4]
- Educator evaluation

Each work group was charged with developing three to five recommendations for consideration by the Department and other education policymakers. Work groups were asked to work toward consensus[5] in crafting recommendations and to ensure that recommendations met the following criteria:

- Allowable under ESSA (considering both statute and regulation);
- Fair and implementable across all local education agencies (LEAs); and
- Attentive to PDE's vision that "Pennsylvania learners will be prepared for meaningful engagement in postsecondary education, in workforce training, in career pathways, and as responsible, involved citizens."

Work groups also were asked to identify priority trade-offs and implications for each recommendation, document key areas of disagreement or concern, and highlight requests for future PDE consideration.

Over the course of three meetings, each work group generated a set of recommendations in its area. This report describes those recommendations and summarizes available research on the extent to which implementation of the recommendations may lead to improved outcomes for students and educators in the state. **This summary of the research and evidence will ensure that the Department and other education policymakers can evaluate recommendations in**

[3] States may submit plans in March or July 2017.
[4] Note that the educator preparation work group was originally charged with considering educator certification; however, because the group's discussion broadened beyond certification requirements alone, in this report we refer it as "educator preparation."
[5] Defined as recommendations that received support from at least three quarters of work group members in attendance during the final meeting on August 30, 2016.

the context of lessons learned from other states and public education systems and helps form a foundation for additional stakeholder engagement efforts.

This report is organized around the four work group topic areas. In each topic area, we provide background on the relevant work group's charge and describe its recommendations. We then address the following questions for each recommendation:

- To what extent is there research evidence that implementation of the recommendation can result in improved outcomes for students, teachers, or schools?[6]
  - Where available research does not provide clear evidence for the policy or practice suggested by the recommendation, to what extent does it align with expert opinion, relevant professional standards, or common state or district policy or practice?

We examined research regarding the following student, teacher, and school outcomes:

- Student-level outcomes: Achievement and academic growth, graduation rates, engagement, attendance.
- Teacher- and principal-level outcomes: Retention, satisfaction, performance.
- School-level outcomes: School climate and culture, student or staff engagement.

We focused on research published in the past 10 years, with the exception of any prominent or seminal research published earlier. Sources comprise peer-reviewed journals; studies released by the U.S. Department of Education, including program evaluations published by the Institute of Education Sciences; and reports from research organizations with rigorous internal review procedures.

In some cases, available research did not match the specifics of the recommendation or addressed only part of the recommendation. In these cases, we present information that is relevant and highlight how it relates to the recommendation.

[6] We do not attempt to characterize the quality or strength of the evidence, merely to summarize what is available.

Assessment

State assessments serve as a barometer of whether students are on track to meet Pennsylvania's academic standards. As required by previous federal law, the current Pennsylvania state assessment system is composed of a number of assessments. These assessments include the Pennsylvania System of School Assessment (PSSA) administered in English language arts (ELA) and mathematics in grades 3–8 and in science in grades 4 and 8; the Pennsylvania Alternate System of Assessment (PASA) for students with significant cognitive disabilities who cannot meaningfully participate in the PSSA; and the Keystone Exams (end-of-course exams in subjects such as Algebra I, literature, and biology).

ESSA Requirements

With respect to assessment, ESSA's requirements are not significantly different from the requirements under the NCLB Act. Students must continue to participate in a common assessment for all students in each state in grades 3–8 in reading or ELA and mathematics and once in high school in those subjects. Science assessments must also be given at least once in each grade span (grades 3–5, 6–9, and 10–12). States continue to have authority to develop assessments in other content areas. Under ESSA, assessments are to be aligned to "challenging" academic standards in mathematics, English language arts, and science; states are required to align these standards to public postsecondary credit-bearing courses. Assessments must "be of adequate technical quality for each purpose defined within the Act" and likely will be evaluated by a peer review process similar to that used in prior years.

In a departure from NCLB, ESSA allows states to use multiple assessments administered over the course of a year in place of a single summative, end-of-year assessment, so long as a single summative score can be produced. States can also administer computer-adaptive assessments. At the high school level, states may substitute a nationally recognized assessment in place of a state-developed assessment, or allow LEAs to do so. Finally, up to seven states will be permitted to pilot innovative assessment measures including competency-based assessments.

ESSA also allows states to use Title I funds to perform audits of their assessment systems, with the goal of ensuring that systems are coherent and purposeful. States may also implement a cap on aggregate testing time.

Overview of Assessment Work Group Recommendations

Work group members felt that the current assessment system is both time and resource intensive, and delivers results on a timeline that is not optimal for informing instructional decision making. The work group discussed a number of options for decreasing the amount of time spent on testing, including reducing or removing open-ended items, or adopting shorter but more frequent tests (perhaps a fall, winter, and spring test, similar to what some districts already do with commercial assessments) that could provide information on where students begin at the start of the school year and where they end. Work group members also discussed matrix sampling—an approach to test design which yields reduced information at the student level, while preserving coverage of content standards at aggregate levels such as school or district level.

Work group members also discussed the idea of allowing students enrolled in eighth-grade algebra to take the Keystone Algebra I assessment instead of the eighth-grade PSSA assessment (another new option under ESSA). However, ESSA also obligates these students to take another mathematics assessment in high school (possibly the SAT or ACT, or another end-of-course assessment in mathematics, such as geometry or Algebra II, which PDE would need to develop). Because the net result of this flexibility does not reduce testing time, the work group decided not to make a recommendation on this issue.

Ultimately, the Assessment work group made the following recommendations:

- **Recommendation 1.** PDE should reduce ESSA-required, statewide testing time for all students.
- **Recommendation 2.** Pursuant to decreasing the time spent on ESSA-required, statewide assessments, PDE should conduct a study to determine the feasibility of administering assessments at multiple points in time to better inform instruction.
- **Recommendation 3.** PDE should utilize a standards-aligned, state-required multiple choice-only assessment to meet ESSA requirements. PDE should encourage local education agencies (LEAs) to utilize performance-based measures for students to demonstrate progress toward achievement of postsecondary goals.

More detail on the work group conversations can be found in the notes from work group meetings available at the PDE website (see *Appendix B* for details). In the following sections, we briefly summarize the work group's discussion regarding the recommendations and the research evidence for each recommendation.

Recommendation 1. PDE should reduce ESSA-required, statewide testing time for all students.

The work group expressed concern that Pennsylvania's current assessments take too much time to administer. The work group hoped that reducing ESSA-required statewide testing time would increase instructional time and allow for increased emphasis on local assessments to provide useful instructional information for both educators and students.

The work group's other recommendations extend this idea by suggesting two potential mechanisms to reduce required testing time: in Recommendation 2, by considering alternative assessment designs; and in Recommendation 3, by using only a single item type (multiple-choice) in ESSA-required statewide tests.

Context and Current Policy or Practice

Depending on grade level, in 2015, students in grades 3–8 spent between nine and 14 hours per year on statewide assessments. Table 1 shows the estimated total administration time (including student testing and test administrative activities) allocated for PSSA assessments in 2015 according to the PSSA test administration manuals.

Table 1. Time (in Minutes) Allocated for PSSAs (2015)

Subject	Session	Grade					
		3	4	5	6	7	8
ELA	1	85	85	85	85	85	85
	2	70	90	95	95	95	95
	3	75	115	115	115	115	115
	4	70	95	90	90	90	90
Mathematics	1	100	100	100	100	100	100
	2	90	90	90	90	90	90
	3	90	90	90	90	90	90
Science	1		75				80
	2		75				80
Totals	Minutes	580	815	665	665	665	825
	Hours	9.7	13.6	11.1	11.1	11.1	13.8

Source: PSSA Handbook for Assessment Coordinators.

Each end-of-course Keystone assessment takes approximately two hours per student (Pennsylvania Department of Education, 2016a). Assuming a 6.5-hour school day and 180 school days, state testing currently accounts for less than two tenths of 1% (about 0.17%) of a typical high school student's year and up to 1.2% of an elementary or middle school student's year.[7] Although the work group's recommendation refers to decreasing statewide testing time, it is important to note that some of the discussion centered on time and energy spent on testing or testing-related activities that are not required by ESSA or the state, such as test preparation or locally mandated testing.

For national context, a 2015 Council of the Great City Schools report (Hart et al., 2015) found that in large urban districts, students spend, on average, between about six hours and nine hours per year on tests mandated as part of federal accountability. The report also found that students in these districts spent an additional six to 11 hours on district-required formative assessments and two to nine hours on other district-mandated assessments. The report highlights redundancy among these assessments. Similarly, a 2014 report by Teach Plus (Teoh, Coggins, Guan & Hiler, 2014) found that students in urban districts spend, on average, 1.7% of the school year on state and district tests in grades 3 and 7. Despite concerns about time spent on mandated testing, data from the 2009 Program for International Student Assessment (PISA) suggest that U.S. students may not spend more time, on average, than students in other countries on standardized assessments (Schleicher, 2015).

Regardless of the actual time spent by students on statewide assessment, there is certainly a perception among some stakeholders that students are being over-tested. For example, a recent survey conducted by the testing firm Northwest Evaluation Association and Gallup (Making Assessment Work for All Students, 2016) reports that more than seven in 10 educators (teachers,

[7] Percentages would be slightly higher for students taking Keystone and PSSA Exams (e.g., eighth-grade math and algebra).

administrators, and superintendents) believe that students are over-tested; however, about half of parents believe that students spend the right amount of time, or too little time, on testing. Results from a 2015 Phi Delta Kappa (PDK)/Gallup poll on attitudes toward education show that 64% of a national sample think there is "too much emphasis on standardized tests" in their communities. In a similar vein, the Center for American Progress (Lazarín, 2014) observed that the regularity of test administration under NCLB has created a public perception of over-testing.

The number of students "opting out" of statewide assessments has also increased during the past several years. In New York, for instance, about 20% of students opted out of state assessments in 2015 and 2016. The rationale for opting out, however, seems to be less about the time spent on assessments and more about the use of the results for teacher and school accountability purposes as well as concern about the appropriateness of academic standards themselves (Pizmony-Levy & Green Saraisky, 2016). In Pennsylvania, opt outs have increased over time but still represent a small proportion of students; in 2015, students opting out of statewide assessments represented a little more than 0.5% of total tests taken.

In 2015, the U.S. Department of Education issued a Testing Action Plan, which suggests that "states place a cap on the percentage of instructional time students spend taking required statewide standardized assessments to ensure that no child spends more than 2 percent of her classroom time taking these tests" (U.S. Department of Education, 2015). The Testing Action Plan further offers examples of high-quality and coherent assessment systems and actions that some states and districts are taking to reduce testing time or otherwise improve assessment systems.

Relevant Research

There is little direct research on the relationship between time spent on testing and student outcomes, though there has been research on the effects of testing and test-based accountability policies more broadly.

In one study that looks directly at the amount of time spent on testing and outcomes, the Council of the Great City Schools (Hart et al., 2015) investigated assessment practices in urban public school districts but found no relationship between the amount of mandated testing time and students' scores on the National Assessment of Educational Progress (NAEP). Nonetheless, the report concludes that decreasing testing time might not have negative consequences either and advises districts to consider carefully the content and quality of tests to reduce redundancy and to ensure that tests achieve their intended purposes without unnecessary burden.

Allensworth, Correa, and Ponisciak (2008) examined a related question: test preparation activities and the impact on Chicago Public Schools students' ACT performance. The authors found that about 60% of 11th-grade English teachers and 40% of mathematics and science teachers were spending at least one month of instructional time on test preparation. The researchers found no evidence that scores improved as a result of the time students spent learning testing strategies or practicing test questions (outside of taking a full, timed practice test). In fact, the researchers noted that, in some cases, increased emphasis on test preparation and test preparation materials tended to reduce ACT scores. The authors noted that previous research has produced similar findings (Briggs, 2001; Scholes & Lain, 1997). Important to note, work group

members explicitly acknowledged this issue, describing how normal instruction in some schools or districts may stop and be replaced by test preparation.

More broadly, a large body of research has examined how test-based accountability policies affect instruction. Although these studies do not speak directly to time spent on testing, they do describe how perceived importance of testing may affect schools and teachers' behavior in terms of changes in curriculum content and emphasis, changes in how teachers allocate time and resources to different instructional activities, and changes in how teachers interact with students (Faxon-Mills, Hamilton, Rudnick, & Stecher, 2013). Some of these changes may be desirable, and some may not. For example, studies document how schools may change the amounts of time spent on subjects (e.g., Rentner et al., 2006; Hannaway, 2007; Amrein & Berliner, 2012; Nichols & Berliner, 2005) or the sequence of content provided in order to emphasize tested subjects or content. Hamilton et al. (2007) and Stecher et al. (2008) describe increases in time spent on ELA and mathematics from 2004 to 2006, during the heart of NCLB implementation. Other studies point to narrowing of the curriculum even within content areas. Many of these same studies also point to other responses, such as focusing on low-achieving students or students near proficiency cut points (Hamilton et al., 2007; Stecher et al., 2008).

Overall, there is little evidence to suggest that reducing time spent on statewide assessments, alone, will lead to improved outcomes for students. Although students can benefit from increased instructional time, research on these benefits generally focuses on significant increases in instructional time from extending the school year, school hours, or afterschool time. It is not clear that time saved by decreasing required testing would be used for instruction, nor that the amount of time saved (likely a few hours) would be sufficient to have any effects.

At the same time, it is likely that districts are spending time not reflected in state test administration windows on activities related to state testing, such as test preparation. Reducing these activities and decreasing the emphasis on the use of test results within high-stakes accountability policies have the potential to influence instruction and student outcomes.

Recommendation 2. Pursuant to decreasing the time spent on ESSA-required, statewide assessments, PDE should conduct a study to determine the feasibility of administering assessments at multiple points in time to better inform instruction.

As previously described, under ESSA, states have the option of using "multiple, statewide interim assessments during the course of the academic year that result in a single summative score" in order to measure students' annual achievement and growth. During discussion on this issue, work group reactions fell into two main categories. Some work group members felt that multiple, shorter testing windows might reduce pressure associated with seeing mandated testing as a single, high-stakes proposition, while delivering valuable, interim feedback. Other work group members noted that, whatever the form, assessments used for federal accountability come with strict administration and security requirements and that multiple administrations would simply multiple these requirements and present new challenges for overburdened staff. A cross-cutting concern was that multiple administrations could obligate districts to follow a particular curricular sequence or, if each assessment covered a broad range of standards, to assess students on some standards they had not yet been taught. Given these tensions, the work group was unable to reach consensus on whether PDE should pursue this option. Instead, the work group

agreed that additional information—including analysis of impacts on schools and districts statewide—would be useful.

Context and Current Policy or Practice

Pennsylvania's current statewide assessment system does not include testing at multiple points in time during the school year.

One state (New Hampshire) currently is conducting a pilot to implement a competency-based system that includes testing at multiple points in time. The Performance Assessment of Competency Education (PACE) pilot permits a small number of selected districts to give the statewide assessment once during each grade span (elementary, middle, and secondary) instead of requiring it annually in grades 3 through 8 and once in high school. For the remaining grades and subjects, districts administer locally developed performance assessments on an ongoing basis. These assessments are intended to be used for local grading purposes and ultimately for accountability purposes as well, although it is not yet clear if this goal will be achieved.

Another example of an effort to test at multiple points in time for accountability purposes is the Partnership for Assessment of Readiness for College and Careers (PARCC) original "through-course" design. This design planned for tests to be given after teachers completed one quarter, one half, three quarters, and 90% of instruction. Some tests were to be in the form of essays and performance tasks, and others were to be quick-turnaround, computer-based exams. All four required formats were to be combined into one end-of-year summative score, which states would use for accountability required by NCLB. However, states engaged in the PARCC consortium ultimately decided not to implement such a design due to concerns about cost, burden, and curriculum alignment. However, a 2011 conference hosted by Educational Testing Service (ETS) on through-course summative assessments described some of the statistical and other challenges in implementing such a design but ultimately concluded that many were surmountable (Jerald, Doorey, & Forgione, 2011).

Relevant Research

Perie, Marion, and Gong (2007) offer a framework to define three major types of assessments[8]:

- **Summative assessments:** Administered one time at the end of a semester or year to evaluate students' performance against a defined set of content standards; often used as part of an accountability program or to otherwise inform policy.
- **Interim assessments:** Serve a variety of purposes, including predicting a student's ability to succeed on a large-scale summative assessment, evaluating a particular educational program or pedagogy, or diagnosing gaps in a student's learning. Such assessments may be given to provide information to teachers or to inform decisions at the school or district level. Typically, interim assessments are given several times a year.
- **Formative assessment:** Used in the classroom by teachers for the explicit purposes of diagnosing where students are in their learning, identifying where gaps in knowledge and

[8] The Center on Standards and Assessment Implementation (CSAI) offers a similar framework (see http://www.csai-online.org/sites/default/files/resources/4666/CSAI_AssessmentTypes.pdf).

> understanding exist, and determining how to help teachers and students improve. Formative assessment is embedded within the learning activity and linked directly to the current unit of instruction. Such assessment may last no more than five seconds and is often called *"minute-by-minute" assessment* or *formative instruction.* Providing corrective feedback, modifying instruction to improve the student's understanding, or indicating areas of further instruction are essential aspects of a classroom formative assessment. Use of formative assessment information generally does not extend beyond the classroom level. The Council of Chief State School Officers (CCSSO) and the American Educational Research Association (AERA) later adopted this definition of formative assessment: "Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes" (McManus, 2008, p. 3).

There is a significant body of well-known and often-cited research supporting the use of ongoing, formative assessment to inform instruction and improve student learning. Black and Wiliam's (1998) seminal work on the impacts of formative assessment on learning is the best example (see also Heritage, 2011; Harrison, 2005; Stiggins, 2008). It is important to note that this research generally regards formative assessment not as a test or instrument administered a few times a year, as the new law permits, but rather as an approach to improve student learning that is incorporated in day-to-day instruction (Heritage, 2011) and one that may be challenging to implement well (see, for example, Marshal and Drummond [2006] on challenges in implementing "assessment for learning").

In addition, there is research on the impacts of "interim" assessments, such as the Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP) assessment, mClass, and Acuity commercial programs. In general, these studies have found that the use of interim assessments has not led to increased student learning, though the findings at times have been mixed. For example, Carlson, Borman and Robinson (2011) found some positive effects in mathematics but not reading, while Henderson et al. (2007, 2008) found no significant differences in gains in mathematics achievement between schools that used quarterly benchmark exams and schools that did not. Cordray, Pion, Brandt, Molefe, and Toby (2012) found no statistically significant impact of the use of the MAP program, and teachers using MAP were not more likely than control group teachers to have applied differentiated instructional practices in their classes. Konstantopoulos, Li, Miller, and van der Ploeg (2016) found some inconsistent evidence that use of interim assessment may be related to improvements for lower achieving students.

Goertz, Nabors, Olah, and Riggan (2009) studied how teachers use interim assessment information and concluded that the success of interim assessments is heavily dependent on the guidance and supports provided by districts and schools to the teachers implementing these assessments, as well as on teacher capacity itself. Shepard (2010) similarly noted that successful use of interim assessment data relies on teacher knowledge and practices, and that it may lead to simple re-teaching—that is, not changing how something is taught but rather simply identifying who or what needed re-teaching. And in a study of interim assessment practices in Philadelphia, Blanc et al. (2010) described the necessary leadership and instructional communities that can make interim assessment data use effective but noted that such communities generally were lacking.

To the extent that PDE wishes to investigate the possibilities of administering assessments at multiple points in time to better inform instruction, current research seems to suggest that such an approach, although technically feasible and perhaps desirable from the standpoint of providing more frequent information, may necessitate significant investments in professional learning along with investments in development of such a system itself.

Recommendation 3. PDE should utilize a standards-aligned, state-required multiple choice-only assessment to meet ESSA requirements. PDE should encourage local education agencies (LEAs) to utilize performance-based measures for students to demonstrate progress toward achievement of postsecondary goals.

The work group's third recommendation suggests that the state assessment, for purposes of compliance with ESSA, be limited to multiple-choice questions only. This recommendation was driven by members' desire to reduce testing time, to increase the speed of receiving results, and to increase the emphasis on locally developed rather than state-prescribed assessment. To that end, the work group's recommendation also suggests that LEAs be encouraged to conduct their own assessments, which could target more in-depth skills in creative ways, such as through performance-based assessments.

An implicit assumption was that, by using multiple-choice questions only, results from the state assessment could be scored quickly and reliably and therefore be returned quickly to students, parents, and educators. There was also a suggestion that a multiple-choice-only assessment might save money currently spent on hand-scoring of open-ended items. The work group did not discuss other types of assessment item formats (e.g., short-answer responses, grid or matching types of questions) that conceivably could be quickly machine-scored and that potentially could have additional benefits.

Some work group members noted that a multiple-choice-only approach might decrease the ability of the state assessment to provide information on student learning with respect to certain state content standards.

Context and Current Policy or Practice

Pennsylvania's current assessment system includes both multiple-choice and open-ended items. Current test design in mathematics calls for 60 multiple-choice questions and three open-ended items; in ELA, the test contains 42 to 47 multiple-choice items and two or three constructed-response items, depending on the grade level; and in science, the test contains 58 multiple-choice and five open-ended items. The Keystone Exams contain about 60% to 70% multiple-choice and 30% to 40% constructed-response items, depending on the specific exam. In addition, as several work group members observed, this recommendation would necessitate a policy change, because Chapter 4 of the Pennsylvania Code requires state assessments to have open-ended items.

Relevant Research

Broadly, the work group's recommendation can be interpreted as calling for a comprehensive assessment system that includes different types of assessments, administered at different levels (e.g., state, district, classroom) for different purposes, all of which are based on an understanding

of these purposes and on agreed-upon standards. Coherent and aligned systems with clear understanding of purpose often have been described as ideal in policy and research literature (e.g., National Research Council, 2001).

More specifically related to the recommendation about multiple choice-only assessments, many policy organizations call for providing a variety of item types in assessments (see, for example, CCSSO's 2014 Criteria for High-Quality Assessments), noting that multiple-choice items may not be adequate to fully assess cognitive performance (Tatsuoka, 1991). Darling-Hammond et al. (2013), citing the Gordon Commission on the Future of Assessment in Education, note that assessments must advance competencies that are reflective of 21st century needs and demands:

> Contemporary students must be able to evaluate the validity and relevance of disparate pieces of information and draw conclusions from them. They need to use what they know to make conjectures and seek evidence to test them, come up with new ideas, and contribute productively to their networks, whether on the job or in their communities. As the world grows increasingly complex and interconnected, people need to be able to recognize patterns, make comparisons, resolve contradictions, and understand causes and effects. (p.3)

Assessing these types of competencies may require items beyond multiple-choice types. In addition, the possibility that guessing and use of test-taking strategies might affect scoring has been cited as a reason to go beyond multiple-choice items (Pollack, Rock, & Jenkins, 1992).

In addition, calls to go beyond multiple-choice questions are likely related to a recognition that the format of tests used for high-stakes purposes—even if other assessments serve other purposes—does send signals to schools about the types of knowledge and skills that matter. Indeed, research suggests that different test formats can influence instruction differently. That is, some studies suggest that the format of an assessment may influence how teachers teach (e.g., Ehren & Star, 2013; Cimbricz, 2002; Resnick, 2006; Webb, 1997 as cited in Herman, 2010), with multiple-choice tests increasing emphasis on basic skills or on specific test-taking skills. The use of direct instruction strategies and an emphasis on whole-class instruction and worksheets also have been linked to tests that emphasize facts and basic skills. However, others note that it cannot be assumed that performance assessments necessarily measure complex thinking skills (e.g. Linn, 1991), and certain studies have demonstrated comparability between multiple choice and short answer items given that similar reasoning procedures are used by students to approach these questions (Huntley et al., 2009; Katz et al., 1996; Lukhele et al., 1993; Shepard, 2008).

Multiple-choice assessments also can offer benefits. From a purely logistical standpoint, multiple-choice tests are readily available and easy to score (Marsh & Cantor, 2014). Conversely, constructed-response items (e.g., short answer) can be costly and time-consuming for both the test taker and the grader (Lukhele, Thissen, & Wainer, 1993). For example, adjusted to current dollars, cost estimates from several National Conference on Next Generation Assessment Systems studies on development and scoring of assessments that include substantial performance components have ranged from about $45 to $55 per pupil, based on the practices used in the early efforts undertaken in the United States (for a review, see Picus et al., 2010). This estimate compares to about $20 per pupil for a largely multiple-choice test. Across studies, costs between performance-based and selected-response tests were generally consistent (Darling-Hammond and Pecheone, 2010).

Multiple-choice assessments may also provide students with opportunities to practice retrieving knowledge from memory, which has been shown to promote the retention of information (McDermott et al., 2014; Roediger & Butler, 2011 as cited in Marsh & Cantor, 2014). For example, when deciding the best answer to select for a multiple-choice question, students must recall information from previous experiences (Marsh & Cantor, 2014).

With respect to performance-based assessments, cost, comparability, and ability to generalize to a broad set of content standards are oft-cited concerns, no matter if they are administered at a state or local level. Evidence about the use of performance-based assessments shows potentially positive influences, such as encouraging teachers to use rubrics in their own instruction, or to otherwise change their teaching strategies to reflect these types of assessment demands (see Faxon-Mills et al., 2013, for a summary). Indeed, in the late 1980s and 1990s, portfolio-based assessments or assessments with performance tasks were not uncommon as part of statewide assessment programs. Lane, Park, and Stone (2002) examined the Maryland School Performance Assessment Program (MSPAP) and found that principals and teachers reported that it contributed to changes in instruction. Stecher, Barron, Kaganoff, and Goodwin (1998), in a study of the Kentucky Instructional Results Information Service (KIRIS), found some positive relationships between assessment results and instructional practices, though the results were inconsistent across grades and subjects.

Evidence accumulated over the last 15 to 20 years suggests that although using a multiple-choice-only assessment may offer financial benefits, it may also carry risks in terms of potentially influencing undesirable instructional or other practices in schools and districts. Performance-based assessment, although holding promise in terms of its relationship to instruction, may also require significant investments to ensure that it can be implemented successfully.

Accountability

Performance-based school accountability systems, which became prominent in the 1990s, are based on a straightforward theory of action: that measuring academic performance and coupling it to rewards and sanctions will cause schools and the individuals who work in them to perform at higher levels (Elmore & Fuhrman, 2001). More explicitly, O'Day & Bitter (2003) explain that accountability systems are comprised of assessments to measure academic outcomes, performance targets to document progress, public reports to provide information to stakeholders, and a system of interventions to stimulate change in classrooms and schools. Thus, in an ideal scenario, an accountability system should ensure that appropriate and achievable goals for school and district educational success are set, and that appropriate resources and supports are in place to respond if the goals are not met **or** if they are exceeded (Ladd, 1996). ESSA challenges states to consider comprehensive measures of school quality and student success, including, for the first time, nonacademic measures, and to act on them through school improvement efforts. The Accountability work group discussed opportunities to develop a variety of measures for use in accountability, as well as how to structure state responses to those indicators.

Ultimately, the Accountability work group made the following recommendations:

- **Recommendation 1.** The accountability system should start with a student-centered approach which considers the whole student experience including academics, physical and cultural environment, and supports
- **Recommendation 2.** The PA accountability system should be based on an array of indicators of student experiences and growth toward college and career readiness, appropriately selected and weighted to serve different purposes, including:
  - Identifying schools for ESSA supports, intervention, and recognition;
  - Timely reporting of meaningful information to schools, policymakers, and communities; and
  - Setting statewide, school, and community goals and interim targets.
- **Recommendation 3.** The PA accountability system will enable system wide, continuous, and sustainable improvement by providing transparent, timely, and meaningful feedback to all stakeholders.
- **Recommendation 4.** The interventions in Pennsylvania's accountability system are evidence-based and applied in ways that are flexible and responsive to varying needs of students, communities, and schools to support the growth of every child. Pennsylvania's system includes a framework for district differentiated recognition, accountability, and assistance. The level of state response is dependent on the tier status of the LEA. The tiered system classifies schools and LEAs on multiple levels based on multiple measures. The level or tier indicates the amount and type of support/intervention needed to improve student outcomes.

In the following sections, we briefly summarize the work group's discussion regarding the recommendations and the research evidence for each recommendation.

Recommendation 1. The accountability system should start with a student-centered approach which considers the whole student experience including academics, physical and cultural environment, and supports.

The work group expressed an overarching intention to establish an accountability system that is more student-centered than prior approaches. Work group members felt that the current system lacks a holistic approach, noting that it fails to account for ways in which race/ethnicity and other student characteristics could lead to unequal opportunities and outcomes. The work group supported the inclusion of multiple data points, such as socioeconomic status, at the individual level (not just the school level) to help account for the challenges faced by children growing up in difficult environments outside schools. Noting that "academics, environment, and supports are inseparable," the group emphasized an approach that considers the academic factors that support readiness to learn for students of all backgrounds, such as equitable access to a rich curriculum, and nonacademic[9] support factors, including for social-emotional wellness and strengthening of school climate or other conditions for learning.

ESSA Requirements

ESSA introduces requirements related to a well-rounded education that expand on NCLB's focus on core academic content. Specifically, ESSA includes the following requirements:

- State, LEA, and schoolwide Title I plans must describe provisions for well-rounded education[10] to all students, particularly for those subgroups of students who are otherwise underrepresented in respective courses, to include "STEM + Computer Science," the humanities, career and technical education (CTE), and accelerated coursework (Advanced Placement, International Baccalaureate, early college, etc.).
- Title I plans must also describe supports for nonacademic factors for readiness, such as parent and family engagement, bullying and harassment prevention, mental health supports, tiered behavioral intervention models, drug abuse and violence prevention, awareness and prevention of sexual abuse, dropout counseling and prevention programs, or training for educators on conflict resolution techniques.

These requirements suggest a desire for all students to be afforded the opportunity to explore rich curricula and experience conditions for learning that address diverse and sometimes challenging student backgrounds.[11]

---

[9] Nonacademic factors include inputs to education other than curricula and other content, with an emphasis on student supports and school climate factors that contribute to noncognitive skills that in turn support learning.
[10] ESSA does not prescribe the content of a well-rounded education but provides numerous examples of courses beyond mathematics and ELA for potential inclusion, including technology and digital literacy, foreign languages, civics, government, geography, philosophy, economics, computer science, engineering, music and other arts, financial literacy, career and technical education, health and nutrition, and physical education.
[11] A well-rounded education and conditions for learning are explicitly supported by the following grants: Supporting Effective Instruction (Sec. 2101), 21st Century Community Learning Centers (Sec. 4201), Student Support and Academic Enrichment Grants (Sec. 4101), and Promise Neighborhoods and Full-Service Community Schools (Sec. 4621).

In addition to Title I plan requirements, ESSA requires the differentiation and identification of schools for supports[12] based on at least four indicators:

- **Percentage of students proficient** in mathematics and reading/ELA on statewide tests;
- **Four-year adjusted cohort graduation rate**, with the option to measure the extended-year rate (high schools);
- **Achievement growth** from year-to-year (on statewide assessment results) or another academic indicator (elementary and middle schools);
- **Percentage of students making progress toward English language proficiency**[13] (ELP); and
- At least one additional indicator of **school quality or student success**

The additional indicator(s) of school quality or student success must be statistically valid, reliable, comparable, and used statewide across schools (though indicators may vary by gradespan). Such factors must also lead to meaningful differentiation between schools, and must be able to be disaggregated by ethnicity and status, such as economically disadvantaged, English learner, and special education. Although there is no requirement that the indicator be nonacademic, ESSA encourages states to measure various factors related to supporting the whole student, including student engagement and school climate and safety.

Beyond the indicators used to identify schools in need of comprehensive or targeted supports and improvement, states must also publicly report additional measures that can help inform schools' nonacademic support efforts, including rates of chronic absenteeism, suspension, expulsion, arrests, violence, and bullying; and achievement and graduation rates of homeless youth, youth in foster care, and military-connected students. These types of data complement accountability indicators and could be collected and studied for future consideration as accountability indicators.

Context and Current Policy or Practice

In 2013, Pennsylvania applied for and was granted flexibility from the U.S. Department of Education regarding some of the requirements of the previous authorization of the Elementary and Secondary Education Act, or NCLB. Under this flexibility waiver, Pennsylvania identified Title I schools as priority, focus, or reward schools on the basis of four Annual Measureable Objectives (AMOs):

- **Test Participation Rate** of 95 percent on the Pennsylvania System of School Assessment (PSSAs) and Keystone Exams in all student groups.

[12] Identification of schools for Comprehensive and Targeted Support Status is based on performance across all or some of these measures (see Recommendation 4).

[13] ELP progress under ESSA is calculated in a manner similar to the "Progress" Annual Measurable Achievement Objective (i.e., AMAO 1) under Title III of NCLB, with the explicit exception that the former may take into consideration such student characteristics as time in language instruction, grade level, age, native language proficiency, and so on.

- **Graduation Rate or Attendance Rate:**
  - High schools: The school must achieve an 85 percent graduation rate or meet the target of a reduction in the difference between its previous year's graduation rate. A Title I school with a graduation rate below 60 percent and not otherwise designated as a Priority school will be designated as a Focus school.
  - Other schools: an attendance rate of 90 percent or an improvement from the previous year is required
- **Closing the Achievement Gap for All Students**. The achievement gap is determined by comparing the percent of students who are proficient or advanced in the baseline year with 100 percent proficiency, and presumes closing 50 percent of that gap over a six-year period.
- **Closing the Achievement Gap of Historically Underperforming Students** –Using the same approach for closing the achievement gap for all students, this AMO applies to a non-duplicated count of students with disabilities, economically disadvantaged students, and English language learners who have been enrolled for a full academic year and take the PSSA, PASA, or Keystone Exams.

As Pennsylvania plans for ESSA implementation, the ESEA flexibility designations remain in effect, as PDE has opted to freeze the Priority, Focus, and Reward school designations through the 2016-17 school year.

In addition to the designations of Title I schools under ESEA flexibility, Pennsylvania measures and reports the performance of all public schools through the School Performance Profile (SPP). The SPP is based on an overall score of 0 to 100, aggregated across multiple weighted measures. These measures focus on academic outcomes such as mathematics, ELA, and science performance. In fact, 90% of the overall high school SPP index score is based on performance on standardized state tests. One proxy for a measure of nonacademic readiness in the SPP index is attendance rate, though it is incorporated as a schoolwide measure, which can mask important data regarding chronic individual student absenteeism.[14]

In addition, Pennsylvania publicly reports other measures such as dropout rates, percentages of classes taught by highly qualified teachers, and enrollment demographics (e.g., percentage of students enrolled in schools by ethnicity, sex, gifted status, economically disadvantaged status, English language learner status, and disability status).

Relevant Research

The work group's recommendation about an approach that takes into account a variety of supports as well as learning environments and experiences can be thought of as encompassing two main ideas: (1) Access to curricula that goes beyond core academic subjects; and (2) Nonacademic school supports.

[14] Chronic absenteeism, in contrast to attendance rate, tracks and flags those students who are absent from school a predesignated number of days (usually 10% of school days) and, therefore, is considered a more actionable indicator.

Access to Curricula Beyond Core Academic Subjects

The beneficial impacts of broad curricula on general achievement (e.g., positive impacts of participation in art courses on literacy outcomes) are supported by numerous evaluations and studies (Powell, 2015; Burton, Horotwitz, Abeles, 1999; Gorard, Siddiqui, & See, 2015; Centers for Disease Control and Prevention, 2010). And a number of researchers have discussed the importance of designing assessment and accountability systems that take a more balanced approach to measuring student outcomes alongside equitable access to these inputs (Tooley & Bornfreund, 2014; Conley & Darling-Hammond, 2013).

A number of states already include measures of student outcomes for subjects beyond mathematics and ELA in accountability or reporting systems. Twenty-nine states currently use achievement indicators in science and/or social studies; of those states, 17 only use a science indicator (Martin, Sargrad, & Batel, 2016). Almost all (49) states include science results in state report cards, and 21 include results of social studies tests in report cards (Nayar, 2015).

A well-rounded education may include many courses beyond these content areas. In a recent national poll, nine out of 10 American adults agreed that the arts are part of a well-rounded education. In the same poll, just 45% of adults agreed that "everyone has equal access to the arts in their community" (Americans for the Arts, 2016). Two states (Connecticut and New Hampshire) currently include participation in the arts as an accountability indicator (Martin, Sargrad, & Batel, 2016).

Five states report enrollment or credit attainment in the arts or foreign languages (Nayar, 2015). Georgia includes the percentage of graduates earning three or more credits in the same world language as an accountability indicator (Martin, Sargrad, & Batel, 2016). Virginia holds schools accountable for the number of students earning credits in courses such as world languages and physics (Virginia Department of Education, 2015).

The benefits of physical education programming on broader learning have been well-demonstrated across various studies as well (U.S. Department of Health and Human Services, 2010). Connecticut holds schools accountable for the percentage of students meeting or exceeding physical fitness standards, and Virginia does the same for participation in a nutrition and physical activity program (Martin, Sargrad, & Batel, 2016).

Other reported measures also can support access to a broad curriculum. Kentucky, for example, reports which schools conduct program reviews of arts and other humanities programs.

Nonacademic Supports

The positive impacts of nonacademic school supports or improvements to school climate and culture on student readiness to learn are well established by a body of research (see, for example: Thapa, Cohen, Higgins-D'Allesandro, & Guffy, 2012; Voight, Austin, & Hanson, 2013; Loukas, 2007).[15] The effects of holding schools or districts accountable for providing such supports through an accountability system is less understood: No systematic research exists on the extent

[15] The work group discussion also addressed a number of specific supports, including trauma-informed social-emotional health supports, nutrition, and physical safety and well-being.

to which using incentives or sanctions in the context of these supports improves outcomes for students, educators, or schools.

ESSA explicitly encourages states to use at least one additional indicator of school quality or student success to measure student engagement or conditions for learning. Chronic absenteeism, usually defined as being absent at least 10 percent or more of the school year, is one such potential measure cited specifically in ESSA; it has been shown in numerous studies to relate adversely to achievement in later grades (Hein, Smerdon & Sambolt, 2013). Indeed, chronic absenteeism is one of the *earliest* research-based warning flags of future academic difficulties (Chang & Romero, 2008) for which direct interventions can be implemented. For example, the New York City Department of Education recently evaluated a mentorship program designed to address chronic absenteeism and found that schools receiving supports were more likely to reduce chronic absenteeism and that students exiting chronic absenteeism showed significant achievement gains (Balfanz & Byrnes, 2013).[16] Five states currently include chronic absenteeism in their accountability measures (Martin, Sargrad, & Batel, 2016).

Indicators of school climate and supports related to positive learning environments are additional options for accountability indicators. In Illinois, schools receive bonus points within their accountability score for "fostering a positive learning environment." Georgia schools receive bonus points for offering programming aimed at improving school climate, such as conflict mediation, mentoring, and positive behavioral interventions (Martin, Sargrad, & Batel, 2016). Texas may allow districts to use district-specific school climate measures in local accountability indicators (Texas Educational Agency, 2016). None of these current measures, however, require disaggregation at the student level, or in the case of the Texas example, are used statewide, and therefore, would not meet ESSA requirements.

Social-emotional learning and supports designed to foster noncognitive skills are other areas of nonacademic supports. These measures show promise in some important regards (Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011), and multiple, publicly-available assessments of social-emotional learning and intrapersonal skills have been independently found to be valid and reliable when deployed consistent with their intended purposes and in appropriate environments (Denham, Ji & Hamre, 2011; Haggerty, Elgin & Woolley, 2011; Humphrey, Kalambouka, Wigelsworth, Lendrum, Deighton & Wolpert, 2011). For instance, in 2015–16, one consortium of California school districts implemented a student survey aimed at measuring nonacademic factors such as "grit." This effort is one initial attempt to measure these factors as part of an accountability system on a large scale—though the results and the consideration of their impact are not yet fully known. Some discussion in the work group reflected doubt about whether it would be appropriate to undertake similar efforts in Pennsylvania given the current lack of evidence.

Nonacademic supports may be measured through surveys of student or educator perceptions, which may raise questions about students' abilities to accurately provide this type of information. However, the accuracy of student perceptions has been supported in the context of teacher evaluation. In one large-scale study, student perceptions, as measured by surveys, of teacher quality were found to have a stronger relationship to achievement scores on statewide exams than

[16] This evaluation studied the impacts of an intensive outreach program to chronically absent students and their families across 100 schools with high chronic absenteeism rates, compared to 46 control schools.

were classroom observations of teachers by other adults (Measures of Effective Teaching Project, 2012). The accumulation of similar, positive findings (Aleamoni, 1999; Follman, 1992, 1995) has led prominent researchers in the field to endorse the use of student surveys, when well designed and well administered, in teacher evaluation (Goe, Bell, & Little, 2008; Marzano & Toth, 2013).

Still, the potential for corruption of survey results in high-stakes situations has led other researchers to stress the importance of improving these tools before use in accountability or differentiation between schools. Numerous technical reasons have been cited for exercising caution—the most challenging being the potential for reference bias, whereby students in schools that emphasize social-emotional health, for example, might apply higher standards to rating social-emotional inputs or status. In one rigorous study, students making higher achievement score gains in oversubscribed charter schools than students in open enrollment public schools were nonetheless more likely to rate themselves *lower* on measures of conscientiousness, self-control, and grit (West, et al., 2015). Angela Duckworth, a prominent contributor to the study of student "grit," has advised against using such measures in high-stakes accountability, citing reference bias, the potential for students to fake data or for teachers to manipulate students into marking the "right answer" and the potential for superficial "parroting" by students of growth mind-set ideas (Duckworth & Yeager, 2015).

There are a few current examples of the use of these types of measures in accountability. New Mexico, for instance, embedded the results of its Opportunity to Learn survey (which questions students about the predominant instructional practices in their classrooms) in its most recent accountability results (New Mexico Public Education Department, 2016). And as previously noted, a consortium of nine California districts, known as CORE, began the use of a self-reported measure of four noncognitive skills[17] for accountability purposes in the 2015–16 school year. A field test, outside high-stakes accountability, demonstrated that the measures showed strong reliability and were positively correlated with key indicators of academic performance, both across and within schools (West, 2016).

In summary, research supports the link between inputs that support the whole child—broad curricula and nonacademic supports—and academic success, but there is little evidence of their efficacy in the context of identifying schools for accountability purposes or designing interventions to support students or schools.

[17] These noncognitive skills are Growth Mindset (belief that abilities can grow with effort), Self-Efficacy (belief in one's own ability to achieve a goal), Self-Management (regulation of one's emotions, thoughts, and behaviors in different situations), and Social Awareness (ability to assume the perspectives of and empathize with others) (CORE Districts, 2016).

Recommendation 2. The Pennsylvania accountability system should be based on an array of indicators of student experiences and growth toward college and career readiness, appropriately selected and weighted to serve different purposes, including:

- ***Identifying schools for ESSA supports, intervention, and recognition;***
- ***Timely reporting of meaningful information to schools, policymakers, and communities; and***
- ***Setting statewide, school, and community goals and interim targets.***

The work group discussed the importance of adopting a multiple-measure system of accountability that, consistent with Recommendation 1, promotes a broader conception of student success and thoughtfully considers the number and weighting of measures, with the greatest emphasis on an accurate measure of academic growth. There was uncertainty that the current growth measure (Education Value-Added Assessment System, or EVAAS—known as PVAAS in Pennsylvania) is "accurate" and "fair" for students of all backgrounds, particularly those with different socioeconomic backgrounds.

An overarching view was that accountability indicators should be selected based on what purposes they serve in the context of reporting, triggering interventions, and serving as the basis for setting long-term goals. Work group members also noted that the indicators should recognize and incentivize the performance of students at all ability levels, without unintended incentives to "push out" certain students. Indicators also should acknowledge the importance of multiple pathways to college and career readiness.

Additional discussion focused on potential measures to use for the indicator of school quality or student success required under ESSA. One reoccurring theme was the potential to use this indicator to broaden the vision for postsecondary placement with greater consideration for placement in career and technical education (CTE) pathways.

The work group also noted that school-level performance targets should be integrated into accountability, such that schools performing at an average level overall would be held accountable for making progress.

ESSA Requirements

Within a state, either all public schools or all Title I schools must currently operate under a state's accountability rules. As described earlier, ESSA requires that states differentiate between schools using (1) proficiency, (2) graduation rate (for high schools), (3) growth or another academic indicator (for middle schools and elementary schools), (4) progress in attaining English language proficiency, and (5) at least one additional indicator of school quality or student success. ESSA further stipulates that the aggregate weighting of the first four indicators is "much greater" than that of the fifth indicator. In addition to being valid, reliable, statewide, and comparable, all accountability indicators must be able to be disaggregated by all racial/ethnic and other disadvantaged subgroups.

ESSA also requires that interim and long-term targets be set by states for proficiency, graduation rate, and English language proficiency progress. Methodologies for target-setting are not prescribed; however, targets must close achievement gaps between subgroups, and if states set extended-year cohort graduation rate targets, those targets must be higher than the goals for a four-year cohort graduation rate.

Context and Current Policy or Practice

The four AMOs described earlier—and approved as part of Pennsylvania's ESEA flexibility application—currently constitute the basis of Pennsylania's accountability system, and are the designations through which PDE associates consequences for school performance.

However, the accountability work group focused discussion on Pennsylvania's SPP index for differentiating schools, which includes a variety of indicators (see Table 2). These indicators are further defined below.

Table 2. Pennsylvania SPP Index (High School)

Indicator Category	Weighting	Sub-indicators/Calculations
Academic Achievement Growth	40%	Percentage of all students meeting annual growth expectations in mathematics, ELA/Literature, and science
Academic Achievement Status	40%	• Percentage of students proficient on statewide tests for mathematics, ELA/Literature, and science (30%) • Percentage progress toward annual proficiency targets in mathematics, ELA, and science for all students (5%) (i.e., "indicator of gap closure") • Percentage progress toward annual proficiency targets in mathematics, ELA, and science for historically underperforming students (5%) (i.e., "indicator of gap closure")
College and Career Ready Indicators	15%	Percentage of students: • scoring at least "competent" on industry-based competency tests (5%) • attaining college and career ready benchmark (5%) • AP/IB, College Credit (2.5%) • PSAT/Plan Participation (2.5%)
Graduation Rate + Attendance Rate	5%	• Graduation rate (HS) or promotion rate (2.5%) • Attendance rate (2.5%)
SUBTOTAL	100%	
Extra Credit for Advanced Achievement	0–6%	Percentage of students scoring *advanced* on core subject statewide exams or industry competency tests, or 3 or higher on AP exams
TOTAL POSSIBLE SCORE	106%	

Achievement status metrics include the percentage of students scoring at least *proficient* in mathematics, ELA/Literature, and science (30 percent) and separate metrics for the percentage

by which the distance between the previous year's proficiency score and the current year's proficiency target is closed (10 percent).

The Pennsylvania Value-Added Assessment System (PVAAS) growth score (40 percent) is a measure of the year-to-year improvement in statewide test performance at the student level, relative to other students with similar test score histories (see the following sections for a more in-depth discussion).

Graduation rate (high schools) or promotion rate (elementary and middle schools) constitutes 2.5% of the SPP index. The graduation rate includes the four-year adjusted cohort graduation rate only, without consideration for students who graduate within five or more years.[18] Attendance rate is weighted at 2.5% as well. The SPP also reports the dropout rate but does not include it in the index score calculation.

A number of college and career indicators aggregate to 15% of the SPP index score. Most of the sub-indicators focus on student *performance level*: scoring at the "competent" performance level on industry-based competency exams, scoring at the college and career readiness benchmark on statewide exams, or obtaining AP/IB or other college credit. Bonus points may be awarded for students scoring at advanced levels on these assessments. One sub-indicator included in the index recognizes *participation* in the PSAT/Plan. The number of advanced placement courses offered by a school is also reported in the SPP outside the index.

Pennsylvania uses *n*-sizes of 30 for the ESEA flexibility designations, and an *n*-size of 11 for reporting the various components of the SPP.

Relevant Research

Because state accountability measures must be applied to all public schools (or all Title I schools) statewide, experimental studies of the impact of certain indicators or weights are rendered problematic. Another major challenge is controlling for factors outside accountability policies, which also vary widely across states (Hanushek & Raymond, 2005 Kane & Staiger, 2002). For example, new statewide assessments might be introduced concurrently with a new accountability system, making it difficult to determine whether subsequent changes in student outcomes are attributable to the assessments, the accountability system, or some other factor. As a result, much of the research concerning the design of accountability systems is correlational.

Since the 1990s, researchers have stressed the importance of using multiple measures in school-level accountability systems, and since that time, state systems have sought to incorporate broader definitions of success (Darling-Hammond, Wilhoit, & Pittenger, 2014; Linn, 2006; Baker, Linn, Herman & Koretz, 2002; Conley & Darling-Hammond, 2013; Martin, Sargrad & Batel, 2016; Kane & Staiger, 2002). Some reasons cited include meeting a diverse array of policy goals, allowing for broader conceptions of school success, maintaining a balanced set of incentives for schools, and increasing the reliability and validity of summative accountability scores (Schwartz, Hamilton, Stecher, & Steele, 2011).

[18] Federal school identification methodologies, now suspended during the transition to accountability under ESSA, used extended-year graduation rates.

In his examination of retrospective data in Ohio, Chester (2005) found that combining the results of multiple measures to determine an overall school accountability score increased the reliability of performance classifications over what those classifications would have been if they were determined by individual accountability measures. Hill and DePascale (2002) similarly found that school-level averages of test scores might have higher levels of reliability than the individual scores that are used to calculate them.

State-level movements toward multiple measures was driven, in part, by negative consequences of NCLB policy, which required a relatively narrow set of measures for accountability (proficiency in ELA and mathematics, graduation rate or another academic indicator and statewide test participation rates), including a narrowing of instructional and curricular focus (Koretz, 2009; Amrein-Beardsley, 2009) without proportional gains on the NAEP (Chudowsky & Chudowsky, 2010; Reback, Rockoff, & Schwartz, 2011; Wong, Cook, & Steiner, 2009). To date, only three of the 50 states and the District of Columbia use fewer than five accountability indicators. Twenty-four states use at least 11 unique indicators in their statewide accountability systems; one state uses 26 indicators. Nationwide, 60 distinct measures are used in accountability systems (Martin, Sargrad & Batel, 2016).

Beyond the consensus regarding the desirability of multiple measures for accountability, there are no acknowledged best practices regarding the precise combination of accountability measures is optimal or how they should be weighted or otherwise combined.[19] Absent research in this area, accountability experts tend to stress balance between complementary measures, such as achievement-growth and achievement-status or course performance and course participation.

In selecting specific indicators, a primary consideration should be the intended use of the measure. Schwartz et al., (2011) identified potential purposes as reporting or monitoring, signaling the importance of a measure to educators, diagnosing and prescribing, and triggering accountability consequences. Particular measures may be more suited for one role or another based on technical validity, transparency, or other characteristics. These mediating characteristics can be considered in terms of key trade-offs identified across the accountability literature which reinforce the notion that accountability indicators tend to be neither "good" nor "bad" but, rather, are better suited for some purposes than for others (Martin, Sargrad & Batel, 2016; Schwartz, Hamilton, Stecher & Steele, 2011). The mediating characteristics include:

- **Breadth versus focus.** As noted, there are multiple reasons for expanding the number of measures included in an accountability system. Simply including a measure in accountability for any reason can signal its importance to educators. Although expanded breadth then would appear to incentivize supports for a broader definition of student success, educators' focus may be too diffused across many measures to make a substantial impact in any one area (Schwartz, Hamilton, Stecher & Steele, 2011). This diffusion must be balanced with the risk of generating a myopic focus similar to that catalyzed under NCLB (Booher-Jennings, 2005; Hamilton et al., 2007).

[19] Weighted accountability systems are otherwise known as compensatory systems. Some states combine multiple measures in conjunctive systems that do not use weightings, but rather identify schools based on measure-level criteria. Virginia, for example, assigns benchmarks for each accountability indicator, and summative ratings are determined based on which indicators meet these benchmarks.

- **Complexity versus transparency.** A number of positive attributes may be associated with increased complexity of an indicator. For example, a value-added measure (VAM) that, through statistical calculations, seeks to control for a student's historical test results or socioeconomic background can be more valid than simpler measures of growth and thus fairer in some regards to certain stakeholders who are impacted by these results. Complexity, however, may hinder the ability of educators or parents to interpret and respond to results and, in the long term, reduce the overall support for the use of the measure (Braun, Chudowsky, & Koenig, 2010).
- **Signaling versus preventing corruption.** As noted, using a wide range of indicators can, in turn, signal the importance of a more balanced approach by educators. The introduction of stakes, however, can present the risk of score manipulation to generate better results. Such cases of cheating, usually through manipulation of standardized test scores, have been systematically documented (Goodnough, 1999; Wilgoren, 2001; Jacob & Levitt, 2002). Some measures are more susceptible to corruption than others—*performance* measures might, for example, discourage educators from increasing the participation of lower achievers. In this case, then, the use of a participation measure can provide balance to the performance measure.
- **Actionable versus summative purposes.** Summative measures are important for tracking school success in achieving overall goals, but they do not always clearly signal what actions should be taken by stakeholders. School-level statistics regarding successful matriculation into college provide an important summative snapshot of ultimate intended student outcomes. These data alone, however, do not fulfill a diagnostic or prescriptive purpose. The introduction of a research-based predictor of future college readiness, such as early literacy (as used, for example, in Ohio's statewide accountability system), provides data that flag the root cause of negative summative results.

Summary of measures by type

In this section, we summarize the relevant research on different types of indicators that have precedent in other states' accountability systems. These include measures of achievement status, achievement targets, achievement growth, achievement gaps, graduation rates, English language proficiency, and additional indicators of school quality or student success.

Achievement Status Measures

Despite widespread professional consensus that achievement status (i.e., a "snapshot" of current-year results), by itself, does not provide a legitimate foundation for the differentiation of schools (Meyers, 2000; Raudenbush, 2004; Linn, 2006), it is a required measure under ESSA, and an important measure, in tandem with growth, for various reasons cited by Linn (2008). First, it is important that all stakeholders understand the absolute achievement level of a school, to support resource allocation decisions and to provide parents and other external stakeholders with objective information for comparing schools and districts. Second, it is important to hold schools to an objective standard (i.e., grade-level proficiency) for achievement aligned with academic

content standards, so that schools cannot consistently rank among the higher performing schools without making real progress toward or beyond grade-level proficiency in core subjects.[20]

The traditional achievement measure of percentage of students scoring *proficient* or higher does not recognize achievement that is approaching proficiency or that exceeds proficiency. Some states use alternate achievement status measures that better recognize all ability levels. South Carolina, for example, uses average assessment scale scores in its accountability system, complemented by conversion tables tied to grade-level proficiency. Ohio uses a proficiency index that places the highest incremental point value on moving students from the Basic to the Proficient performance level and otherwise recognizes all ability levels. It is unclear whether ESSA will permit states to use such alternate achievement status measures in statewide accountability systems.

Achievement Targets

Setting appropriate schoolwide targets for long-term student performance is an inexact science with scarce supporting research or analysis. NCLB's Adequate Yearly Progress (AYP) goals were considered unrealistic for many schools (Linn, 2005; McCombs, Kirby, Barney, Darilek & Magee, 2005).[21] In 2003–04, a quarter of the nation's schools failed AYP as defined by their states. The percentage of schools that *made* AYP in 2003–04 ranged from 95 percent of schools in Wisconsin to 23 percent of schools in Alabama and Florida. (Le Floch et al, 2007).

Responding to widespread consensus that NCLB's AYP goals were unrealistic (Linn, 2005; McCombs, Kirby, Barney, Darilek, & Magee, 2005), under ESEA flexibility, the U.S. Department of Education provided states with three options: (a) cut the number of nonproficient students in half within six years; (b) reach 100% proficiency by the year 2020; or (c) propose their own "ambitious, but achievable" goals provided that subgroups further behind made greater progress. The "cut-in-half" option was largely based on a statistical analysis of 10 states' performance data conducted by the Education Trust (Ushomirsky, Hall, & Haycock, 2011). Almost half of states receiving flexibility chose to halve the number of nonproficient students, and only two chose to retain goals toward 100% proficiency (Hall, 2013). To date, there exists no published analysis of states' ability to meet either short- or long-term ESEA Flexibility targets.

The cut-in-half option was not without some element of arbitrariness. In fact, the Education Trust also recommended at the time that, after the states' transition to more rigorous college- and career-ready standards (required by the second year of ESEA flexibility implementation), long-term targets be adjusted downward to reflect the performance of the top 10% of schools in a given state. This recommendation was in line with Linn's recommendation (2006) to base targets on the performance of the top 10% to 20% of schools. Under ESEA flexibility, Colorado and Wisconsin set targets according to this recommendation. Research on best practices in setting short-term (e.g., annual) targets is scarce; however, Perie, Park, and Klau (2007, CCSSO) rightly point out that as proficiency approaches 100%, the more difficult it becomes to close

[20] In order to hold students to a meaningful objective state standard, it is also important for a state to set meaningful cut scores for assessment performance levels.

[21] Under NCLB, all public schools were required to set annual goals towards 100 percent proficiency by the school year 2013-14. Goals could escalate annually in equal increments or in a step-wise fashion that escalated every 2-3 years.

achievement gaps completely; incremental approaches that call for equal gains over each annual or interim period, then, might not adequately consider this factor.

Another critical issue for states to address is an apparent misalignment between measures of progress and accountability designations. Under ESEA Flexibility, a number of states, including Colorado, Massachusetts, and Minnesota, developed school accountability indices that embed progress against targets within their school performance indices. Most states, however, do not integrate this mechanism into their systems, and, as a result, schools may consistently attain average or high accountability scores without actually making absolute progress. This issue can be particularly important for students in lower achieving subgroups. As Sargrad, Marchitello, and Hanna (2015) point out, the highest performing schools for all students often demonstrate the largest achievement gaps.

Achievement Growth

On one hand, proficiency measures hold all students to the same grade-level expectations. Linn (2008), however, points out that proficiency scores assume that student progress is wholly attributable to schools and teachers and that students enter the school year at the same achievement starting point. Growth scores, in contrast, take into account the starting point of the student and may also consider factors outside a school's control.

Forty-six states measure growth in ELA and mathematics for accountability, and seven states also measure growth in science and social studies (Martin, Sargrad & Batel, 2016, p. 13). For some states, growth measures can account for up to 50% of an overall accountability score. Though not part of the formal accountability system, Pennsylvania's SPP does include a measure of academic growth (accounting for 40% of the score).

Growth measures may be divided into three classifications: "simple" growth models that consider only the actual or *absolute* growth of each student independent of other factors; models that calculate the "value added" by a teacher or school to a student's growth by considering additional factors, such as prior test scores and student characteristics; and student growth percentiles (SGPs), which account for testing histories but typically not student background (Goldhaber & Theobald, 2013).

An example of a "simple" growth model is Florida's "learning gains" approach (Florida Department of Education, 2016). In this method, students who are less than proficient must advance a certain amount within their current performance level to be credited for growth. For students who already meet the "passing" level (Level 3 or 4), any increase in scale score is recognized, while students at the upper level of achievement (Level 5) need only maintain their current performance level to attain a positive growth score. Although the learning gains measure does not control for factors outside a school's or teacher's control, the advantage of this and other simple growth models is that it is easy to understand and provides straightforward growth expectations across students, classrooms, and schools.

Both value-added growth measures (VAMs) and student growth percentile (SGP) models are typically derived from regression analyses. Both types of measures are relative, or normative, measures of growth that compare students with similar testing histories. Some VAMs go beyond

the EVAAS by also controlling for student background characteristics such as race or socioeconomic status. Research shows that VAMs with controls for both student background and past test results increased the rankings of teachers of disadvantaged classrooms compared to VAMs controlling only for test histories (Goldhaber & Theobald, 2013).

Because of their statistical complexity, these types of models may be less readily understandable than simple growth models. One study in Pennsylvania found general awareness and understanding of value-added models (VAMs) to be "quite low" (Dembosky et al., 2005).

Achievement Gap Measures

According to the most recent available data, at least 19 states use some measure of achievement gap closure in statewide accountability systems (Education Commission of the States, 2013), either as an achievement status measure or measure of gap change across school years. Achievement gap measures may be a particularly important guardrail for high-performing schools, which often demonstrate the highest gaps in achievement between low and high performers (Martin, Sargrad, & Batel, 2016).

ESSA does not require the use of an achievement gap accountability indicator, but it does require that subgroups with lower achievement scores set appropriately aggressive targets that result in the closure of achievement gaps. Pending federal rules would require that failure to meet interim (e.g., annual) achievement goals for any particular subgroup over a number of years be reflected in accountability indicator measures.

Pennsylvania's current "gap closure" measure is a measure of performance against a state-set proficiency target. We did not identify any research related specifically to the design of achievement gap measures.

Graduation Rate

Beginning with the class of 2010-11, the U.S. Department of Education (ED) required all states to publicly report the number of students attaining a standard diploma within four years, with the option to also report extended-year graduation rates. ESSA stipulates that the four-year adjusted cohort rate be used for accountability purposes and provides the option of also using an extended-year rate.

States' accountability practices around graduation rates have tended to go beyond federal requirements. Although every state except Washington embeds the four-year graduation rate into statewide accountability (Washington uses the five-year rate exclusively), 37 states also use a five-year or longer graduation rate. Colorado allows schools to use the highest of the four-, five-, six-, or seven-year rate.

Although graduation rates may be important indicators, they do not capture the attainment of other high school benchmarks, such as advanced diplomas and general education diplomas. In Texas, which has a state college and career readiness standards definition that embeds Algebra II performance, accountability points are awarded to schools graduating "distinguished" students who, among other requirements, must score a 3.0 or higher on the Advanced Placement mathematics test. Virginia employs a graduation index that recognizes general education diplomas and certificates of program completion and embeds dropout rates.

Although graduation rates are important cumulative measures to include in accountability, dropout rates can provide valuable "real-time" data regarding the performance of a particular cohort. Dropout rates are included in the systems of 11 states. Massachusetts and Texas also measure the rate at which schools reengage dropouts (Martin, Sargrad & Batel, 2016).

Despite a record high 82% nationwide graduation rate for the 2013-14 school year, disaggregated subgroup rates for that school year lagged considerably: 76% for Hispanics, 73% for African Americans, and 70% for American Indian/Alaska Natives. The overall graduation rate, then, can mask subgroup persistence levels. States such as Washington have responded by including graduation rate gap measures in overall accountability systems.

Although ESSA does not require the use of subgroup graduation rate measures for identifying comprehensive support schools or "consistently underperforming" schools for targeted support, it does require that the "additional targeted support" schools, identified by the lowest performing subgroups across all federal accountability indicators, consider this measure. Under currently proposed rules, states have flexibility in defining how "consistently underperforming" schools are identified based on subgroup performance—this would be an additional policy vehicle for prioritizing subgroup graduation rate accountability (see proposed rules in the Federal Register).

Graduation Rate Target-Setting

Hall (2013) emphasizes the need for ambitious but achievable target-setting for graduation rates, including for all students and all subgroups. He notes that without meaningful graduation goals, there can be pressure to raise achievement by pushing lower performers out of school. Ushomirsky, Hall, and Haycock (2011) recommend setting long-term targets to reduce by half, over six years, the difference between a school's current four-year graduation rate and 90% or, for extended-year cohort rates, the difference between current rates and 95%. These goals, like the recommended achievement targets, are based on the performance of high-performing schools.

English Language Proficiency

A major shift under ESSA is the requirement to embed progress toward English language proficiency (ELP) into the statewide accountability system. Under proposed rules, this measure may now consider student characteristics such as language instruction, grade level, age, and native language proficiency, all of which can impact how long it takes to attain proficiency (Collier, 1995; Thomas & Collier, 1997). States may ostensibly continue to define progress toward English language proficiency in a flexible manner.

Currently, six states embed some form of ELP performance in their accountability systems (Arizona, Colorado, Georgia, Massachusetts, Texas, and Illinois). Three states specifically use progress toward proficiency attainment.

Additional Indicator of School Quality or Student Success

States continue to position college and career readiness at the center of overall educational strategies. At least 21 states have adopted definitions of college and career readiness at the student level, and 42 states mandate the use of individualized learning plans implemented at the school level to help students stay on track to postsecondary success (Mishkind, 2014). However,

only 60 percent of students who began a four-year undergraduate degree in 2008 finished within six years; rates are lower for African-American (41 percent) and Hispanic students (54 percent) (U.S. Department of Education, 2016). Regarding graduates on vocational pathways, employers and researchers continue to note the lack of academic and other employability skills (e.g., interpersonal skills) (Dymnicki, Sambolt, & Kidron, 2013; National Network of Business and Industry Associations, 2014). Although Recommendation 1 noted the opportunity for states to use the additional accountability indicator(s) to measure nonacademic readiness, it might also be used to promote college and career readiness (noting that the two are not mutually exclusive).

Measures that promote college and career readiness may be placed into two categories: cumulative measures of success, such as the rate of college enrollment; and future readiness indicators that are collected earlier in a student's educational trajectory in order to inform timely instructional and support responses.[22] Many states include both types of measures in accountability. Connecticut, for example, includes high school CTE participation as well as college enrollment. Hein, Smerdon, and Sambolt (2013), in a survey of correlational research, identified a range of research-based predictors of future readiness that states may consider for use in accountability systems.

An inventory of states' college and career readiness indicators within accountability systems is detailed below (see Table 3).

Table 3. College and Career Readiness Indicators Used in Statewide Accountability Systems

State Accountability Indicators	Number of States Using for Accountability
Participation or performance in college entry exams (SAT, ACT, ACCUPLACER, or COMPASS)	24
Participation or performance in advanced coursework (AP, IB, or dual/concurrent enrollment)	22
Performance or participation in CTE courses or other experiential programming	26
Postsecondary enrollment[23]	4

Despite the consensus among policymakers and researchers on the importance of using multiple measures in accountability systems, there is little research to support decisions regarding which combinations of measures more meaningfully identify schools than others. Regardless, it is clear that using multiple measures can increase the validity and reliability of overall accountability determinations and support a richer theory of action for identifying leverage points for school improvement. Policymakers should consider the trade-offs between transparency, accuracy, fairness, and potential for corruption as they consider each of the required indicators under ESSA. Fairness in particular must be considered from the perspective of each of the relevant stakeholders.

[22] In this regard there can be an overlapping of the purposes of accountability systems and early warning systems. States should select accountability indicators in consideration of how the two systems complement each other.
[23] ESSA now requires that states publicly report college enrollment rates within one year of high school graduation, where data are available. In addition, other postsecondary indicators in use by states include military enrollment (Missouri) and percentage of graduates not requiring college remediation (Georgia).

Recommendation 3. The Pennsylvania accountability system will enable system wide continuous and sustainable improvement by providing transparent, timely, and meaningful feedback to all stakeholders.

The work group addressed issues related to the communication, presentation, and use of accountability results, and expressed the importance of providing transparent, timely, and meaningful feedback to all stakeholders. Conversation followed two distinct threads: (1) the aggregation, formatting, and presentation of data in meaningful and transparent ways, and (2) practices that support data-driven decision making.

Members discussed whether to attach an accountability designation to schools beyond a numeric score, such as an A-F grade. Some members stressed caution in making such a policy change, noting that such designations may be meaningful to some stakeholders and not to others. Other members suggested that a system of five performance tiers, each driving specific levels of support, might be worth considering. Overall, members felt that report cards should balance diagnostic and summative information and data. Members also noted that socioeconomic context and progress toward five-year goals are important considerations in transmitting this information. Furthermore, work group members felt that stakeholders should be able to discern quickly what the information and data reported to them mean.

Other discussions addressed various practices related to developing a culture of data-driven decision making, particularly in order to inform instruction. Timeliness of data results was flagged as an important issue. Some data, such as attendance, are available quarterly, and other data, such as state test results, are not available until the following September—these factors should be considered in the theory of action around using data. Other participants noted the extent to which data might be "misused" and pointed to the importance of educator capacity-building around analyzing and using data to inform instruction. The use of professional learning communities (PLCs) to review data collaboratively and the deployment of technology-based tools to share information were also flagged as potential strategies for supporting data-driven action.

ESSA Requirements

ESSA requires the identification of the lowest performing 5% of all public schools across all accountability indicators for comprehensive support status. Proposed rules would require that an overall performance level be assigned to each public school based on all accountability indicators (out of at least three overall performance levels), and that individual accountability indicator scores are each assigned one of at least three performance levels.

ESSA also requires that LEAs annually publish report cards for each school that disclose accountability performance results along with other, federally-prescribed data. ESSA-related supports in this area include:

- The use of State Assessment Grants under Title I to design annual report cards in an "accessible, user-friendly manner";
- Opportunities to fund evidence-based, data-driven decision making strategies for school improvement under Title I;

- Expanded allowable uses of professional development funds under Title II for the training of all educators in data analysis and using data results to drive instruction; and
- Additional opportunities under Title IV to fund technology infrastructure projects that support data-driven decision making initiatives.

Context and Current Policy or Practice

Pennsylvania uses a report card that aggregates performance metrics into a score between 1 and 100. This score does not result in the assignment of an overall performance level, such as an A–F grade. Federal accountability categories (Priority and Focus schools) are currently the only overall designations that trigger formal consequences for schools. Report cards are used to inform supports but otherwise do not formally signal particular supports or rewards—the meaningfulness, therefore, of report card results in terms of specific consequences may be interpreted differently by different districts, schools, and educators.

Pennsylvania has not implemented a formal system of professional learning communities, though some districts may have adopted related practices, and the state is making online classes and other voluntary supports available through www.pdesas.org.

Relevant Research

One of the primary roles of school-level accountability systems is motivation (Herman, Baker, & Linn, 2004). Through a combination of public reporting and administration of consequences, accountability systems intend to galvanize stakeholder behaviors to improve student outcomes. Numerous researchers have affirmed that accountability systems can result in greater overall school performance (Hanushek & Raymond, 2006; Deming, Cohodes, Jennings, & Jencks, 2016). One theory of action posits that to achieve desired student outcomes, performance-based accountability systems must "focus attention on goals, enhance the availability of valid information on which to base decisions, increase motivation to strive for goals and build and focus capacity to improve" (Boyle, Taylor, Hurlburt, and Soga, 2010). This vision relies on information that is symbolically powerful, understandable, and translatable into decisions and actions (Herman et al., 2004; Polikoff et al., 2014).

Identifying the final format and systems for reporting results in a transparent, timely, and accessible way is essential for meaningful action by stakeholders (Howe & Murray, 2015; Mikulecky & Christie, 2014; Baker, Linn, Herman, & Koretz, 2002; Karene and Staiger, 2002). The research-based literature regarding successfully leveraging accountability results can be demarcated into two categories: 1. the types of final measures, formats, and presentation styles that support intuitive data interpretation and use by stakeholders for their intended purposes; and 2. the development of adequate systems, supports, and conditions to support the leveraging of data for the intended purposes.

Presentation of Data

The Data Quality Campaign (2016) points out that planning related to the meaningful deployment of performance data can begin by first considering the ultimate needs of all stakeholders. This approach emphasizes starting with the purpose of the data in mind—for

example, teachers needing to understand the trajectory of their students' progress as well as having supporting information that can augment their approaches, such as information about student background or support services received.

Considering the diversity of needs, states must give serious consideration to how results will be meaningfully framed for stakeholders. States may aggregate these results into overall school ratings or maintain them at the measure level in dashboard or scorecard form (Brown, Wohlstetter, and Liu, 2008). Summative ratings have been credited with providing a high-level filter for stakeholders who need to sift through sometimes overwhelming amounts of data, and criticized for oversimplifying a school's performance. Data dashboards, designed well, can provide a rich portrait of a school's performance that conveys its strengths and weaknesses in the context of goals, year-to-year performance, and performance relative to other schools; at the same time, it may be unwieldy for purposes of making cross-school comparisons of *overall* performance (Cowart, 2010; Ushomirsky, Williams, & Hall, 2014). Neither approach need be mutually exclusive, however, and states can and do combine these approaches.

States are closely watching the status of a pending ESSA accountability rule that would require that the state school differentiation system "results in a single rating from among at least three distinct rating categories for each school" (Rule 200.18 from U.S. Department of Education, 2016a). States including California have advocated against this requirement, including through the submission of public comments that charge that a single summative rating for schools "necessarily glosses over differences in performance across indicators and inappropriately draws school leaders['], stakeholders['], and the public['s] focus on the single rating rather than a more robust reflection of performance demonstrated by the individual indicators" (California Department of Education, 2016). The state's comments specifically noted that the use of a summative rating could mask disaggregated subgroup performance (California Department of Education, 2016).

The perceived utility of overall summative grades for schools undoubtedly depends on the perspective of the stakeholder. In a report regarding stakeholder perceptions of state report cards gathered through focus groups in South Carolina, parents repeatedly expressed positive regard for high-level differentiation between schools through an overall summative rating, though they did not negatively regard report cards that did not use a summative rating (Southeast Comprehensive Center, 2016). Business owners participating in the same study described how high-level information filters, such as school grades, can efficiently support analyses of geographic levels of workforce readiness. In contrast, a population of academics and practitioners stressed the primacy of measure-level results in order to diagnose root causes of poor academic performance accurately and expressed strong reservations regarding summative rating systems.

Positive claims regarding the clarity of summative ratings originated with the introduction of the first A-F system in Florida in 1999 (Howe & Murray, 2015). The Foundation for Florida's Future supported this innovation with the rationale that assigning a letter grade is a way to report a school's effectiveness in a manner everyone can understand. (Foundation for Florida's Future, 2014).

Other states, including Arizona, Indiana, Utah, and West Virginia, made similar public statements supporting the understanding of accountability results, especially in the context of aiding parent and family engagement (Indiana Department of Education, n.d.). West Virginia's public-facing overview of its A-F system indicates that "this new system relies on an inherent

understanding of what an A through F grade indicates. School grading commands a focus on learning because parents, educators and administrators all understand the meaning of letter grades" (West Virginia Department of Education, 2016).

Still, despite enthusiasm for summative designations, no research base supporting their connection to improved student outcomes exists. By the same token, almost no primary research exists to detract from their use. Critiques of summative ratings, particularly A–F systems, that do exist generally align with claims that they:

- Mask the results of important individual performance measures (by their very nature);
- Are not as easy to understand as proponents claim; and
- Can unfairly brand a school as a low performer despite strong, school-based efforts.

The most widely cited study of the first issue was conducted by Adams and Forsyth (2013) for the Oklahoma Center for Education Policy and the Center for Educational Research and Evaluation. The authors found, based on a retrospective analysis comparing A-F grades to mathematics and reading achievement scores, that "a truly comprehensive [school] evaluation system is best not boiled down to a single value because it masks the very complexity it is trying to capture." One specific finding to support this conclusion is that within-school achievement gaps were highest within A and B schools, and students receiving free and reduced-price lunch in D and F schools had higher average reading, mathematics, and science achievement scores than did students receiving free and reduced-price lunch in A and B schools. This study is not without significant limitations, however. Most important, the Oklahoma accountability system does not embed achievement gap values of students receiving free or reduced-price lunch or any other disaggregated subgroups in its accountability system. In fact, Oklahoma does not embed achievement status of disadvantaged students as a discrete indicator in its system at all. These results may point more clearly, then, to accountability design lacking in strong regard for disaggregated subgroup measures than to systematic masking of subgroup performance.[24]

A second criticism is that the A-F scale itself is not as clear or intuitive as many claim it to be. Although it is widely accepted that A and B ratings are desirable, for example, stakeholders may have diverging opinions as to the meaning of a C grade (Southeast Comprehensive Center, 2016). In other words, although the sequence of values in an A-F system is clear to many stakeholders, the relative value of ratings to each other might not be.

This issue can also be highlighted by comparing the distribution of school ratings across a small sample of states. Table 4 demonstrates how differently states may set intervals between overall school performance levels. Although Oklahoma's distribution of A-F schools resembles a bell curve, under Florida's A-F system, over one third of schools received an A grade. The meaning of an A in each of these systems appears to be different. In addition, rating systems such as Virginia's that prioritize accreditation, a status that denotes a minimum level of satisfactory performance, lead to high percentages of schools at the highest rating (78%). Colorado likewise designates a high percentage (71%) of schools with the highest performance rating. In the

[24] The Oklahoma accountability system weights achievement status of all students at 50%, achievement growth of all students at 25%, and achievement growth of the lowest 25% of performers at 25% (bonus points are awarded for graduation rate and other indicators).

absence of appropriate context, a stakeholder in either of these states might inappropriately assume that a school in the highest performance category is exceptional, particularly given the number of overall performance levels (five) in each case.

Table 4. Distribution of Statewide School Ratings Across All Public Schools in Four States, Based on 2014–15 School Year Performance Data

State and Ratings (highest to lowest)	Percentage of Schools Receiving Lowest Rating	Percentage of Schools Receiving Intermediate Rating (not highest or lowest)	Percentage of Schools Receiving Highest Rating
Oklahoma A, B, C, D, F	10%	78%	12%
Florida A, B, C, D, F	6%	55%	35%
Virginia Fully Accredited, Approaching, Improving, Warned, Accreditation Denied	1%	18%	78%
Colorado Performance Plan, Improvement Plan, Priority Improvement Plan, Turnaround Plan	3%	26%	71%

Notes: Percentages do not sum to 100% for each state due to additional schools pending final rating determinations; Colorado data based on 2013–14 school year data, pending public release of aggregate statistics for 2014–15 data.

Finally, Howe and Murray (2015) make the claim that A-F systems can unfairly brand schools as poor performers, in spite of strong school-based efforts, which can lead to demoralized staff and an inability to retain and attract talent. This argument is not without merit. Unless key measures specifically control for racial, socioeconomic, and other background characteristics, an overall summative rating can disproportionately reflect the negative impact of out-of-school factors in schools with large disadvantaged populations. State stakeholders have echoed the researchers' sentiments regarding adverse impacts on staffing (Wagner, 2015; Ableidinger, 2015). Similar to the VAM context described earlier, trade-offs relating to fairness, accuracy, and transparency must be considered in the context of the purposes of the accountability system.

Using Accountability Data (Data-Driven Decision Making)

The previous discussion focuses on options for framing or presenting accountability results in a meaningful, transparent way. Generating accountability results in a useful format alone will not ensure, however, that these results will be used, let alone lead to meaningful overall school progress. It is only when states, schools, and districts establish the proper supports and resources for leveraging data within the context of accountability systems that these results lead to positive impacts on student achievement (Schildkamp, Lai & Earl, 2013). These supports and resources are identified in data-driven decision making research literature as being related to data

infrastructure and access, timeliness of data delivery, time and venues for data analysis, and educator capacity building.[25]

The development of a robust data infrastructure including the facilitation of efficient access to student-level data by teachers is identified as a key factor in successful data-driven decision making culture. Faria et al. (2012) found that school-level data infrastructure is related to high student achievement on state tests. This factor also is identified as one of the most challenging barriers for states and districts. Coburn, Honig, and Stein (2005) found that many school districts lack the technical capacity to facilitate efficient access to data for teachers. Two studies including Pennsylvania educators yielded similar findings. Demboski (2005) found, through case studies of six districts and one charter school in southwest Pennsylvania, that districts lacked comprehensive and integrative data systems due to small size and limited resources. In some cases, Intermediate Units (IUs) provided technological and analytical support. In the Implementing Standards-Based Accountability (ISBA) study, which included interviews with state officials and case studies of 18 schools across California, Georgia, and Pennsylvania, one third to one half of mathematics teachers had access to technology systems to support data analysis, but only one third to one half of those teachers found these systems useful (Stecher & Naftel, 2006).

Timeliness in providing data also is cited as an important enabler of data-driven action. Coburn et al. (2005) emphasize the importance of data timeliness and note the frequent mismatch between the fast pace of decision making in schools and actual receipt of supporting results. States, then, might consider the complementary roles of summative and formative assessment results when designing data-driven systems. More than 80% of superintendents in the ISBA study found that results from local tests, not statewide tests, were more useful at the school level, at least partially because spring test results are released too late to be useful for the current student population.

In addition to timely reporting of achievement and accountability information, teachers need time to meaningfully analyze data. However, schools and districts that successfully designate and protect time for teachers to do so over the long term are generally the exception (Feldman & Tung, 2001; Ingram, Seashore, & Schroeder, 2004). One reason cited is that pressures to keep pace with curricula limit teachers' opportunity and willingness to reflect on formative results and then reteach content. One strategy for allocating time is to embed collaborative data practices into the school day that take advantage of regularly scheduled department meetings. Some survey results indicate that collaborative data activities are at least as prolific as individual data analysis (U.S. Department of Education, 2011), and researchers affirm that collaborative data practices can successfully lead to higher student achievement (Faria et al., 2012).

Finally, various studies identify **educator capacity** as an essential catalyst for data-driven decision making and find that educators frequently lack skills and knowledge to formulate questions, identify relevant indicators, interpret results, and develop solutions (Choppin, 2002; Feldman & Tung, 2001; Supovitz & Klein, 2003).

[25] The research included in this section may include some results outside of the context of accountability; the principles discussed here, however, are independent of the accountability context—they are important regardless of the purposes for which the data was generated.

One early study of district administrators documented their difficulty in knowing what type of data to analyze (Kennedy, 1982). Additional capacity challenges have been documented by Khanna, Trousdale, Penuel, and Kell (1999) and Penuel, Kell, Frost, and Khanna (1998). McCaffrey and Hamilton (2007) surveyed teachers in Pennsylvania receiving PVAAS scores and found that only a small minority of them were confident in their ability to translate results into instructional adjustments. Data Quality Campaign, aside from recommending integration of data literacy training into ongoing professional development, has stressed its inclusion in preservice programs. Other researchers emphasize the inherent capacity-building benefits of collaborative data practices (Feldman & Tung, 2001).

As the research described in this section demonstrates, factors impacting *how* the results of accountability systems are communicated are comparably important to *which* accountability indicators are used. Overall results may be transmitted through some combination of summative and dashboard formats that can serve a variety of purposes, including providing a strong symbolic message that motivates stakeholders and detailing strengths and weaknesses for a more informed approach to school improvement. These approaches can balance their respective weaknesses: summative ratings can oversimplify performance and dashboard approaches may not be as user-friendly or motivating. States should ensure that their strategies support data use by considering data infrastructure, timeliness in delivering results, providing time for educators to use data, and building their capacity to do so.

Recommendation 4: The interventions in Pennsylvania's accountability system should be evidence-based and applied in ways that are flexible and responsive to varying needs of students, communities, and schools to support the growth of every child. Pennsylvania's system should include a framework for LEA differentiated recognition, accountability, and assistance; the level of state response is dependent on the tier status of the LEA. The tiered system classifies schools and LEAs on multiple levels, based on multiple measures. The level or tier indicates the amount and type of support/intervention needed to improve student outcomes.

The work group discussion of this recommendation touched on many topics, including tiers of support, evidence-based interventions, and external supports to low-performing schools. First, participants discussed the need for tiered identification of districts coupled with sufficient support for districts to implement evidence-based interventions. Consistent with the principle of tiered support, the work group underscored the need for differentiation associated with district needs. Some work group members advocated for as many as five tiers, based on multiple measures, which would enable Pennsylvania to recognize struggling districts as well as high achievers. Others voiced support for opportunities for districts to move to higher tiers accompanied by mechanisms to celebrate progress.

With regard to the needs assessment and planning process, work group members discussed the tension between flexibility and state oversight—that is, districts should have the flexibility to identify interventions that best meet the needs of their schools, but this ability requires greater agility and expertise on the part of the state when monitoring implementation and progress. Acknowledging the variation in district capacity, some participants supported a more robust role for the state in districts that want or need more direction as they develop their improvement plans.

Work group participants also discussed the need for an infrastructure and financial supports for districts in the more intensive intervention tiers. For example, some participants suggested that some configuration of support teams should provide direct support to schools in need of comprehensive support, specifically with regard to improvement plans and implementation.

Finally, work group members acknowledged the need for a suite of evidence-based interventions and described a role for the state in identifying these interventions. More broadly, work group participants underscored that the accountability system needs to provide timely dissemination of information to drive and promote continuous planning and improvement.

ESSA Requirements

In a departure from No Child Left Behind's (NCLB) prescribed menu of interventions, ESSA empowers states and districts to map out a system of supports and interventions for the lowest performing schools. No longer will states face a mandate to implement a specific set of interventions associated with successively more punitive accountability designations.[26] Rather, states may identify their own interventions and turnaround strategies, as long as those strategies are supported by evidence.

Some general parameters do exist—under ESSA, states and districts share responsibility for supporting low-performing schools, although there is some latitude in the balance of responsibilities. At a minimum, states must ensure that districts conduct a school-level needs assessment for comprehensive support and improvement schools, in partnership with a range of stakeholders. The State Plan must account for student performance against state-determined long-term goals, identify resource inequities, and include evidence-based interventions. The state must also establish statewide exit criteria for schools identified for comprehensive support and improvement. For comprehensive support and improvement schools, districts must decide which "more rigorous" actions must be taken by such school (which may include addressing school-level operations) if there is no improvement within the state-determined number of years. States may initiate additional improvement in districts with large numbers of schools needing improvement and, consistent with state law, establish alternative evidence-based strategies that can be used by districts to assist schools.

Although NCLB mandated the use of "scientifically based" interventions, ESSA adopts a broader focus on "evidence-based" approaches. On September 16, 2016, the U.S. Department of Education released nonregulatory guidance regarding the use of such evidence. The guidance is anchored by a cycle of improvement that includes identifying local needs, selecting evidence-based interventions, planning for implementation, implementation, examination, and reflection. The guidance elaborates on the definition of "evidence-based" interventions found in ESSA, and explains that an intervention may have a rationale based on high-quality research if it is grounded in "a well-specified logic model that is informed by research or an evaluation that suggests how the intervention is likely to improve relevant outcomes" (U.S. Department of Education, 2016a).

[26] Under NCLB, schools in improvement status were variously required to implement school choice, supplemental educational services, new curricula and/or various alternative governance structures (e.g., replacing staff), depending on the number of consecutive years for which AYP targets were not met.

WHAT IS AN "EVIDENCE-BASED" INTERVENTION? (from section 8101(21)(A) of the ESEA)

"the term 'evidence-based,' when used with respect to a state, local educational agency, or school activity, means an activity, strategy, or intervention that –

(i) demonstrates a statistically significant effect on improving student outcomes or other relevant outcomes based on –
  (I) **strong evidence** from at least one well-designed and well-implemented experimental study;
  (II) **moderate evidence** from at least one well-designed and well-implemented quasi-experimental study; or
  (III) **promising evidence** from at least one well-designed and well-implemented correlational study with statistical controls for selection bias; or

(ii)
  (I) demonstrates a **rationale based on high-quality research findings or positive evaluation** that such activity, strategy, or intervention is likely to improve student outcomes or other relevant outcomes; and
  (II) includes ongoing efforts to examine the effects of such activity, strategy, or intervention.

Context and Current Policy or Practice

Throughout the implementation of NCLB, PDE sought to establish and implement a comprehensive statewide system of school support that could address the needs of districts and schools with a history of persistent low performance. When the U.S. Department of Education extended flexibility on some provisions of ESEA, PDE submitted a revised approach to accountability designations and supports. Under ESEA flexibility, Pennsylvania schools designated as Priority schools were required to implement interventions aligned with the turnaround principles outlined by the U.S. Department of Education and could benefit from a wider set of supports. According to PDE documents, these tools included a curriculum audit process, online diagnostic assessments, an early warning system to improve graduation rates, and a kindergarten entry inventory. Pennsylvania schools that failed to exit Priority status within three years were required to implement significant changes aligned to School Improvement Grant (SIG) options (Pennsylvania Department of Education, 2014a). In addition, Priority schools were assigned an academic recovery liaison who supported the development of a comprehensive improvement plan. PDE would then approve the plan and interventions, in consultation with the academic recovery liaison (Pennsylvania Department of Education, 2014b).

Looking toward ESSA implementation, Pennsylvania has the opportunity to revisit existing practices, intensify or augment supports, and establish processes for identifying and selecting evidence-based practices.

Relevant Research

By 1999-2000, at least 20 states described strategies to assist schools that were identified for improvement under the Improving America's Schools Act (IASA; Goertz & Duffy 2000. Beginning in 2002, NCLB accelerated this work, and codified requirements for state support; Section 1117(a)(4) specified that states were required to establish school support teams and to designate and use distinguished teachers and principals to support school improvement, in addition to other state-specific supports. By 2006-07, 42 states, the District of Columbia, and Puerto Rico reported using support teams, and 26 states reported using distinguished principals

and teachers as either a primary support mechanism or as an important component of their support system (Taylor, Stecher, O'Day, Naftel, & Le Floch, 2010). Other mechanisms of support included regional centers and outside consultant groups.

The provision of supports to low-performing schools evolved further under ESEA Flexibility. Le Floch and Tanenbaum (2016) conducted a study of state supports under ESEA Flexibility in 12 states; among the states they studied, they observed (a) increased reliance on regional entities to provide support to schools, (b) increased emphasis on the role of the district in supporting low-performing schools, and (c) frequent use of tiered systems of support that ensure that the most intense interventions are implemented in the lowest performing schools.

Many studies of state supports for low-performing schools are descriptive policy scans, some based entirely on extant policy documents and others incorporating interviews with state officials. For example, staff at Regional Educational Laboratory (REL) Central conducted a 50-state scan of policies for intervening in chronically low-performing schools (Klute, Welp, Yanoski, Mason, & Reale, 2016). Among their key findings are the following:

- Forty-seven states have policies related to interventions in staffing in chronically low-performing schools;
- Thirty-one states have policies related to closing chronically low-performing schools;
- Thirty-seven states have policies related to financial incentives or interventions in chronically low-performing schools; and
- Thirty-two states have policies related to interventions in the day-to-day operation of chronically low-performing schools.

Most studies of state provision of supports to low-performing schools are descriptive, mixed-methods studies that provide case studies or survey results regarding the implementation of the components of state support as well as analyses of student outcomes (see, for example, Becker, Koger, Sinclair, & Thacker, 2009; Huberman, Dunn, Stapleton, & Parrish, 2008; Le Floch et al., 2011). In very broad terms, these studies find that components of the state supports are implemented with fidelity at the local level and that school respondents report moderately positive impressions of the support. However, associated analyses of student outcomes have failed to detect statistically significant effects.

Thus, although there is a descriptive research base on the implementation of state systems of support for low-performing schools, very few studies have established a causal connection between mechanisms of support and improved outcomes. One exception are studies of Massachusetts schools in designated "Commissioner's districts." Schools in Massachusetts receive a numerical accountability rating in which 1 is highest performing and 5 is lowest performing. Districts with schools in Level 4 or 5 also receive these designations; Level 5 districts become "Commissioner's districts" and receive additional support, including:

- **District Liaisons.** Massachusetts state education agency staff members serve as project managers and coordinate support to the districts, overseeing implementation of the state's strategy for school turnaround.

- **Priority Partners.** External partners support turnaround efforts regarding students' social, emotional, and health needs. These partners work to maximize learning time, effective use of data, and district systems of support.
- **School Redesign Grants.** SRGs are competitive funds that support turnaround efforts in persistently underperforming schools. SRGs are provided through federal School Improvement Grant (SIG) funding.

A series of analyses of student outcomes in schools receiving School Redesign Grants (SRGs) demonstrated statistically significant gains among students in the SRG schools compared with students in the comparison schools. The effects were statistically significant after the first, second, and third years of SRG implementation on both the ELA and mathematics assessments (LiCalsi & Piriz, 2016).

Although Massachusetts' efforts appear promising, there are insufficient causal studies to provide a clear roadmap for states seeking to redesign systems of supports. However, over the past decade, scholars and practitioners have attempted to synthesize lessons learned from research and practice. Most recently, education leaders from Colorado, Louisiana, Massachusetts, and Tennessee collaborated with the Center for American Progress to identify state-level turnaround tenets grounded in research and best practice (Martin, Sargrad & Batel, 2016). Some of these tenets are directly applicable to Pennsylvania's current work:

1. Grant districts and ultimately the state the authority to intervene in failing schools;
2. Provide significant resources to support planning and restructuring and leverage competitive grants;
3. Treat the district as the unit of change and hold it accountable for school improvement;
4. Create transparent tiers of intervention and support combined with ongoing capacity building and sharing best practices;
5. Promote stakeholder engagement;
6. Create pipeline programs for developing and supporting effective turnaround school leaders; and
7. Embed evaluation and evidence-based building activities in school implementation.

Educator Preparation

ESSA eliminates the "highly qualified teacher" (HQT) requirements of No Child Left Behind and provides opportunities to identify strategies for recruitment, retention, and support of talented educators. Members of the Educator Preparation work group initially focused on educator certification; however, as the work group recommendations evolved, discussion centered on strategies to prepare, recruit, and retain a diverse educator workforce, which resulted in the following recommendations:

- **Recommendation 1.** The Department should promote and increase opportunities to recruit, retain, and ensure a diverse, talented, and supported educator workforce.
- **Recommendation 2.** The Department will define effective teachers as those who strive to engage all students in learning, demonstrate instructional and subject matter competence, and continuously grow and improve.
- **Recommendation 3.** The Department should promote and support collaborative in-field, practical experiences as a crucial component of educator preparation.
- **Recommendation 4.** The Department should promote and increase opportunities to recruit, retain, and support diverse and talented school leaders.

In the following sections, we briefly summarize the work group's discussions regarding the recommendations and the research evidence for each recommendation.

ESSA Requirements

The quality of the education workforce is critical to student success. No Child Left Behind and the subsequent Race to the Top (RTT) and Elementary and Secondary Education Act (ESEA) waivers strongly encouraged—and in some cases mandated—increased accountability and rigor in talent management policies and practices. Although ESSA has relaxed these requirements in many cases, it also has created an opportunity to leverage and advance what has been learned.

With respect to certification and licensure, ESSA no longer requires the highly qualified teacher status mandated by NCLB and the Individuals with Disabilities Education Act; however, states will be required to develop standards for highly effective educators and must continue to report relevant data. In addition, requirements for equitable access to effective educators for children in high-poverty and high-minority schools remain in place with ESSA, with specific emphasis and recognition on the importance of effective educators and high-quality instruction for these students.

ESSA includes important levers to ensure equitable access, most particularly in the way states and districts attract and retain effective teachers. First, ESSA offers funding flexibility under Title II, Part A to improve educator quality in its recently released Non-Regulatory Guidance for Title II A: Building Systems of Support for Teaching and Leading (U.S. Department of Education, 2016b). In particular, modifications have been made to the funding formula that allows additional funding for states and districts serving a larger proportion of low-income students. This funding can be used to attract, recruit, and retain teachers, including the use of strategies such as differential compensation and providing teacher leadership opportunities. The law also includes the Teacher and School Leader

Incentive program fund, which authorizes competitive grant funds to local education agencies to support human capital management systems. Title II Part A funds can also be used to implement induction and mentoring programs, teacher residencies, and strategies to improve all students' access to effective teachers and to improve working conditions.

Recommendation 1. The Department should promote and increase opportunities to recruit, retain, and ensure a diverse, talented, and supported educator workforce.

One of the most pressing issues facing stakeholders at all levels is how to staff classrooms with teachers who are equipped with both the content knowledge and pedagogy to support a diverse student population. The work group reviewed existing state and national data about teacher recruitment, retention, and preparation practices in an effort to make informed recommendations.

Recurrent teacher shortages can be linked to a decline of candidates entering the profession and to high rates of teacher attrition (Shields et al, 2000). Some specific teaching fields have suffered consistently – for example special education, mathematics, and science – with attrition even worse where poor working conditions and low wages exist (Lieb, et al, 2016). Teacher shortages and issue of retention are costly in terms of resources and time spent to attract new teachers (Podosky, Kini, Bishop, Darling-Hammond, 2016).

Demographic trends – nationally and within Pennsylvania – support the need to both recruit and support all teachers so that they remain in the profession. Recent statistics have shown that the number of teacher certifications issued by the state's Department of Education was down by 62% between 2012 and 2015 (PA Certification by Subject Areas, 2011- 2016).

Likewise, recruitment and retention of minority teachers is needed. A state-by-state analysis conducted by the Center for American Progress in 2014 revealed that 96% percent of teachers in the state of Pennsylvania were white (Boser, 2014). Research supports the need to recruit teachers from historically underrepresented groups in order to match the increasingly diverse student populations. Studies cite benefits of improvements in teacher retention and gains in student achievement (Clotfelter, Ladd, & Vigdor, 2007; Dee, 2004; Egalite, Kisida, Winters, 2015). Likewise, literature suggests that special educators with similar cultural backgrounds can positively impact achievement of students with disabilities and potentially limit the number of inappropriate referrals to special education (Tyler, Yzquierdo, Lopez-Reyna, & Saunders-Flippin 2004). Yet, research conducted in 2002 indicated that only 14 percent of special education teachers are from historically underrepresented groups (Billingsley, 2002).

Therefore, the work group acknowledged that increasing the state's capacity to recruit, retain, and ensure a diverse, talented, and supported educator workforce would reduce the amount of time school districts spend on recruiting effective teachers and, more important, would be a major benefit to students (Tyler, Yzquierdo, Lopez-Reyna, & Saunders-Flippin 2004;Villegas & Irvine, 2010; Egalite, Kisida, & Winters, 2015; Grissom & Redding, 2016). The Educator Preparation work group explored and discussed various strategies in support of this recommendation and made the following four distinct sub-recommendations:

- **Sub-Recommendation 1a:** Promote and market teaching as a valued and respected profession;

- **Sub-Recommendation 1b:** Improve recruitment efforts through the use of financial incentives and by targeting diverse populations;
- **Sub-Recommendation 1c:** Investigate certification requirements considering quality and effect on diversity recruitment; and,
- **Sub-Recommendation 1d:** Strengthen educator support across the career continuum.

Sub-recommendations are described in more detail in the following sections, followed by a summary of relevant research and/or examples of similar current policy or practice.

Sub-Recommendation 1a: *Promote and market teaching as a valued and respected profession.*

Given the data on recruitment rates reviewed by the work group, members brought forward recommendations to strengthen the perception and status of the teaching profession as one method to increase both interest and entrance into the profession through the following strategies:

- Launching a statewide marketing campaign to recruit teachers;
- Increasing opportunities for talented high school students and nontraditional teacher candidates to explore teaching as a career; and
- Improving the public perception of teaching.

June 2014, the Organisation for Economic Cooperation and Development (OECD) completed its *Teaching and Learning International Survey (TALIS)* of high school teachers and found that only 34 percent of the respondents in the U.S. believed teaching is valued by U.S. society (Organisation for Economic Cooperation and Development, 2014). The McKinsey & Company report, titled *Closing the Talent Gap: Attracting and Retaining Top-Third Graduates to Careers in Teaching* (Auguste, Kihn, & Miller, 2010), surveyed approximately 1,600 college students and practicing teachers to determine the types of incentives and policy changes needed to motivate more college students – in particular top performing students – to enter the teaching profession. Findings revealed that respondents perceived the profession as less desirable than other professions, with only 9 percent of college student respondents saying they planned to enter the teaching field. While similar reports exist, there is little research on the effects of efforts to elevate and market the profession.

That said, there are a number of programs designed to both increase the perception of the profession and to recruit potential candidates by exposing and incentivizing high school students through hands-on experiences and financial support, such as the following programs:

South Carolina's Teaching Fellowship program was designed to recruit talented high school seniors into the teaching profession while giving them the tools necessary to become effective teachers. Every year the program offers fellowships to over a hundred high school seniors who have displayed high academic achievement and have exhibited a history of service to their schools and communities. Students selected for a fellowship receive up to $24,000 in scholarships and are offered numerous professional development opportunities (Center for Educator Recruitment, Retention, and Advancement, 2011).

Illinois's Golden Apple Scholars program aims to recruit talented high school seniors who have shown potential to become exceptional teachers. Students who enter the program are given up to $23,000 in tuition support and are provided with paid summer institutes that offer courses and extensive classroom experience. Students who enter the program are specifically trained to teach in the highest need schools (Golden Apple, 2015).

Educators Rising closely resembles a career pathway system in that it provides a recommended program of study, extended learning activities, and guidance on how high school students can transition to postsecondary education and certification. The program focuses on increasing diversity in participants' local teaching workforce, steering high school students toward exploring high-need subject areas, and supporting certified and employed participants to ultimately take on teacher leadership roles. Educator Rising has program affiliates across the nation, including states like Arizona, Delaware, Nebraska, Ohio, and West Virginia.

The ***Teach Campaign,*** funded by Microsoft Corporation and State Farm Insurance and initiated by the U.S. Department of Education, seeks to promote teaching as an attractive, viable career option for current high school students. The campaign advertises through various media to promote teaching as a profession. The campaign also operates a website that helps students navigate information and resources relative to teaching as a career.

Today's Students, Tomorrow's Teachers (TSTT) was founded more than 20 years ago to promote diversity in New York's teaching workforce and is chartered by the New York State Board of Regents. The organization's programs center on school-based mentoring, targeting minority students who are economically challenged to engage their interest in teaching as a career. TSTT offers college tuition assistance (up to 50%) and provides beginning teaching placements for students in their communities. Currently, just under 1,000 New York students participate in TSTT.

Nevada's Clark County School District has implemented a number of recruitment strategies to extend the reach to a wide range of potential teaching candidates. Though Clark County still suffers from teacher shortages, recent recruitment efforts have had success in hiring new teachers (Whitaker, 2016). Part of Clark County's success can be attributed to its creative recruitment strategies, which are used to recruit as widely as possible. Clark County makes use of the Internet to recruit, promote, and advertise on nearly a hundred different websites. In addition, Clark County examines trends in applications to modify recruitment strategies. When Clark County observes an increase in interest from a particular region, it often recruits more heavily in that area. Another unique aspect of Clark County's recruitment strategy is its reliance on current and retired school administrators to recruit candidates rather than using a full-time recruitment staff.

In summary, at this point little research exists on efforts to extend the marketing, recruitment, and perception of and into the teaching profession. While several examples of programs to strengthen field perception and increase recruitment into the profession, concerted effort to evaluate the impact of these efforts are warranted.

Sub-Recommendation 1b: *Improve recruitment efforts through the use of financial incentives and by targeting diverse populations.*

The work group described the following specific strategies:

- Improving compensation
- Using incentives to recruit in high-need areas and to increase the diversity of the workforce

Compensation

Research about teacher compensation continues to suggest that salaries affect the labor market decisions that teachers make (Ingersoll, 2003; Keigher & Cross, 2010; Borman & Dowling, 2008). Both the survey research (which asks teachers whether and why they chose to join or leave the teaching profession) and econometric literature (which reports on observed changes in teacher recruitment or attrition as these relate to teachers' salaries) suggest that salaries matter. Though somewhat mixed, research suggests that although teachers enter the workforce due to an altruistic desire to "serve," perceptions of low compensation have persuaded potential candidates away from the field – most particularly within minority populations (Auguste, Kihn, and Miller, 2010). A report by McKinsey & Company surveyed approximately 1,600 college students and practicing teachers and found that most of the students (top-third) underestimated teacher compensation (Auguste, Kihn, and Miller, 2010. Moreover, on average studies have shown that teacher pay is not competitive compared to many labor markets (Baker, Sciarra, & Farroe, 2015), especially in difficult-to-staff subjects like mathematics and science (Beaudin, 1995). Research also finds that higher salaries would make teaching a more viable career option for math and science majors in college (Milanowski, 2003) and for high-performing college students from the top-third of their college classes (Auguste, Kihn, & Miller, 2010).

In response, states and districts have experimented with compensation and other financial incentives to make teaching more attractive to high-quality candidates and as a means to retain qualified teachers (Balter & Duncombe, 2008; Hirsch, Koppich, & Knapp, 2001; Kolbe & Rice, 2006; Loeb & Miller, 2006; Strunk & Zeehandelaar, 2011). Findings in these studies suggest a mixed level of success in teacher recruitment and retention. Several studies suggest that pay can affect both the short- and long-term supply of teachers (Ladd & Sorenson, 2016) and an area in which a significant majority of teachers believe would positively impact teacher retention (Scholastic and the Bill and Melinda Gates Foundation, 2012). A comprehensive literature review conducted by Guarino, Santibanez, and Daley (2006) suggested that higher salaries were associated with lower attrition. Using nationally representative data from the largest national data set on teacher mobility (the *Teacher Follow-up Survey*), Ingersoll (2003) asked former teachers about the factors that led them to leave and found poor salaries at the top of the list of sources of dissatisfaction both for teachers who changed schools (49%) and for teachers who left the profession (61%). Similarly, an international literature review by Dolton (2006) concluded that improving teacher pay could reduce teacher shortages. Finally, a "meta-analysis" by Borman and Dowling (2008) similarly concluded that salaries are an important factor in the retention of beginning teachers and even more so for experienced teachers.

Studies have also indicated that teachers are more likely to leave in lower paying districts (Adamson and Darling-Hammond, 2012). One study of labor market decisions by teachers concluded that salary increases of at least $18,000 are needed to retain teachers in difficult-to-staff schools; however, salary increases have somewhat limited success when employed in isolation (Feng, 2014). Clotfelter, Glennis, Ladd, and Vigdor (2006) found much smaller incentives to be effective (for STEM teachers).

There is a growing body of research that supports the perception that schools with a larger percentage of students from low income or poverty backgrounds and/or with lower performance on statewide assessments, have larger numbers of un- or under-qualified teachers (Lankford, Loeb & Wyckoff, 2002). Using data from the Schools and Staffing Survey (SASS), the National Center for Education Statistics reported that the percentage of movers and leavers in schools with 75 percent or more of students approved for free and reduced-price lunch was 22 percent compared to 12.8 percent in school with 0-34, respectively (U.S. Department of Education, 2012-2013). Moreover, Pogodzinski (2000) found that low wages – as compared to relative wages in the county – played a significant factor in the use of emergency permits or waivers.

Findings in both Texas and New York suggest that first year teachers who were identified as less effective at improving student test scores have higher attrition rates and that more effective teachers transfer to higher achieving schools, while less effective teachers transfer to lower performing schools (Boyd, Grossman, Lankford, Loeb, & Wyckoff (2008). Goldhaber, Gross and Player (2007) substantiated this claim and found that the teachers who transfer and leave teaching are less effective than those who remain. Salaries also ranked among the top "dislikes" for top-performing "irreplaceable" teachers (TNTP, 2013). Overall, the research suggests that the more effective teachers are likely to either leave the profession or transfer out of a high-need, low performing school and to some extent salary can play a factor in effective teachers remaining in the profession and in high-need schools. However, the research investigating the correlation between teacher salaries and student performance is relatively mixed (Hanuschek & Rivkin, 2004; Figlio, 2002; Loeb & Page, 2000).

There have been some efforts to entice effective teachers to move to high-need, low performing schools. A study of the Talent Transfer Initiative (TTI; Glazerman, Protik, Teh, Bruch, & Max, 2013), which was implemented in 10 districts in seven states, found positive effects of offering transfer incentives. The initiative offered teachers ranking in the top 20% within their subject in terms of student achievement $20,000 in paid installments over a two-year span to transfer to designated low-performing schools. The study found that TTI successfully attracted high value-added teachers, had a positive impact on mathematics and reading test scores, and had a positive impact on teacher retention

Other financial incentives, such as signing bonuses, student loan forgiveness, housing assistance, and tuition reimbursement have all been employed to assist in teacher recruitment and retention (Hirsch, Koppich & Knapp 2001; Feistritzer, 1997). Likewise, funding for these initiatives has taken many forms, such as the No Child Left Behind Act and Teacher Incentive Fund. States and districts have explored the use of housing incentive programs; however, these have not yet been studied to determine the level of efficacy as a recruitment or retention strategy. Opportunities to assume leadership roles have also been associated with improved teacher satisfaction and retention (Berry, Daughtrey & Weider, 2010; Brooker & Glazerman, 2009). Last, given that

strong preparation for teaching increases teacher efficacy and likelihood of staying in the profession (Darling-Hammond, Chung, & Frelow, 2002) loan forgiveness programs have been employed, though they have not been researched to determine their efficacy.

Increasing Diversity of the Workforce

A 2016 report from the U.S. Department of Education notes that there is a disparity in racial diversity in the schoolteacher workforce. In the 2011–12 school year 82% of all public school teachers were white (U.S. Department of Education, 2016d). What is more, a state-by-state analysis conducted by the Center for American Progress in 2014 revealed that 96% percent of teachers in the state of Pennsylvania were white (Boser, 2014). Using data from the Schools and Staffing Survey (SASS) and the Teacher Follow Up Survey (TFS) Ingersoll (2016) illustrated that there continues to be a persistent gap between the percentage of minority students and the percentage of minority teachers in the U.S. school system. For instance, in the 2011-12 school year 44.1 percent of all elementary and secondary students were minority, but only 17.3 percent of all elementary and secondary teachers were minority. Although some studies suggest that recruitment of minority teachers has been effective, the increasing diversity of the student population has offset any gains in recruitment (Ingersoll & May; 2016; U.S. Department of Education, 2016d).

In an effort to increase the diversity of the educator workforce, the Brown Center on Education Policy (2016) describes four areas of concern impacting both the recruitment and retention of a diverse educator workforce:

- A smaller proportion of minority populations earn college degrees.
- There is a lower interest in teaching careers among minorities than whites.
- Minority teachers are hired for teaching jobs at lower rates than white teachers.
- Minority teachers are retained in lower rates than whites (Putman, Hansen, Walsh, & Quintero, 2016).

The Brown Center report further stated that achieving a diverse teacher workforce is a long-term policy goal that requires a plethora of strategies to help minorities succeed in college and encourage them to enter the profession. On this latter point, the report added that improving working conditions in public schools, leadership, and salaries are important first steps toward making teaching more attractive to all (Putman, Hansen, Walsh, & Quintero, 2016).

Goldhaber, Theobald, and Tien (2015) cite multiple research studies that substantiate the need to diversify the teaching profession and offer multiple theoretical arguments for diversifying the teaching profession, including the following:

- Minority students, especially those in underserved schools, benefit from adult role models in a position of authority.
- Minority teachers have higher expectations for minority students.
- Cultural differences between teachers of different backgrounds in terms of instructional strategies could lead to positive impacts (Goldhaber, Theobald, & Tien, 2015).

Additional research on the topic suggests positive impacts on student achievement when minority students are taught by teachers of their own race ethnicity (Clotfelter, Ladd, & Vigdor, 2007; Dee, 2004). One such study examined the data from Tennessee's Project STAR class size experiment and found improved achievement in both mathematics and reading when students were assigned to a teacher of the same race (Dee, 2004).

Using data from the Using the National Center for Education Statistics' Schools and Staffing Survey/Teacher Follow-up Survey, the Learning Policy Institute (LPI) report that while recruitment efforts over the last decade have been successful, data suggest that minority teachers depart from their schools at higher rates – mostly due to working conditions (Ingersoll & May, 2016). This is particularly true in difficult-to-staff schools where minority teachers are more likely to be placed. The researchers also noted that minority teacher retention was correlated to shared decision-making and leadership structures (Ingersoll & May, 2016).

In summary, research shows that compensation can impact teacher retention. There is also research that a diverse teaching force may improve student achievement. However, while recruitment efforts have met some success it has not kept pace with the growing population of diverse students. Moreover, recruitment efforts need to be complemented by retention efforts given the high rate of minority teachers who exit the profession. Some policy makers argue that to improve the diversity of the teaching force, systemic changes will need to be made that may include improving working conditions, leadership and compensation—which is supported to some degree in the research.

Some examples of successful models are included below:

- The Arizona Future Educators Association (FEA), developed a "grow your own program" in 2005, to create a student organization in conjunction with a two-year program for Arizona high school juniors and seniors interested in becoming teachers. The primary purpose of Arizona FEA is to recruit young people to the field of education and to improve retention rates by providing additional hours of classroom teaching experience while still in high school (National Center to Improve Recruitment and Retention of Qualified Personnel for Children with Disabilities, 2012).
- Arizona developed the Paraprofessional Tuition Assistance Grant (PTAG) to support paraprofessionals already residing in high needs districts to become highly qualified speech language pathology assistants (SLPAs) or special education teachers. The purpose of PTAG is to increase retention of highly qualified special education personnel, especially those serving rural schools (National Center to Improve Recruitment and Retention of Qualified Personnel for Children with Disabilities, 2012).
- The Para-to-Teacher District/University Partnership was first developed in 2009, in response to Utah's critical shortage of highly qualified special education personnel. Aware of the data showing high retention rates for paraprofessionals becoming special education teachers (National Center to Improve Recruitment and Retention of Qualified Personnel for Children with Disabilities, 2012).

- Georgia is House Bill 280, which created **differentiated compensation for mathematics and science teachers**, offering different incentives to elementary (K–5) and secondary (6–12) teachers. In its inaugural year, the governor and legislature approved 9.59 million of funding for the program under the Quality Basic Education Program (Georgia Department of Education, 2011). The only cost to districts associated with the salary increases was an increase in certain benefits.
- In Oklahoma, the **Teacher Shortage Employment Incentive Program** provides loan reimbursement (or the cash equivalent) to teachers prepared in Oklahoma who teach mathematics or science in an Oklahoma public secondary school for at least five years. More information about this program can be found at: http://www.okhighered.org/otc/tseip.shtml.
- In Ohio, the **STEM Teacher-Signing Bonus Program**, created in 2009, offers a signing bonus or loan forgiveness to new STEM or foreign language teachers. The bonus is $4,000 per year or the equivalent amount in loan forgiveness (for up to five years, or $20,000).

Sub-Recommendation 1c. *Investigate certification requirements considering quality and effect on diversity in recruitment.*

Pennsylvania currently requires all teacher candidates to take a basic skills, general knowledge, and professional knowledge test prior to being issued a license (Pennsylvania Department of Education, 2016b).

The work group discussed the following strategies in particular:

- Evaluating the impact of the basic skills test for entry into profession on quality of candidates
- Establishing an equity committee that focuses on the successes of and barriers to recruiting a diverse workforce

Research conducted by the University of Washington and the Urban Institute found significant trade-offs when basic skills tests are used as screening devices. Specifically, researchers found that teachers who would have been effective in the classroom are found ineligible due to their poor test scores. Conversely, ineffective teachers with high test scores are allowed in the classroom (Goldhaber, 2006). These tests can especially have disproportionate effects on minority candidates striving to become educators. A study from the Education Testing Service found significant gaps in average scores on general skills tests in reading, writing, and mathematics between test takers of different racial subgroups and whites, with African Americans showing some of the largest disparities (Education Testing Service, 2011).

In summary, research does not support basic skills tests as valid measures to determine teacher effectiveness. Moreover, these basic skills tests have disproportionate effects on minority candidates, potentially creating a barrier to minority populations pursuing teaching as a profession. No research was found to support the strategy of establishing an equity committee.

Sub-Recommendation 1d. *Strengthen educator support across the career continuum.*

The work group discussed the following strategies in particular:

- Strengthening mentoring and induction programs and expectations
- Exploring improved partnerships across educator preparation programs (EPPs) and LEAs
- Leveraging educator evaluation systems to provide high-quality, targeted professional learning

Induction and Mentoring

In order for educators to be as effective as possible, they need to be supported by the state, district, and school in which they teach. Finding effective ways to support all teachers—most particularly new and struggling teachers—is especially important as the nation is experiencing a rise in teacher shortages. Research suggests that induction and mentoring can increase not only teacher retention rates but also demonstrated gains in teacher effectiveness and student learning—though these gains correlate to the quality of support provided (Ingersoll, 2012; Ingersoll & Smith, 2004; Strong, 2006; Villar & Strong, 2007; Kaiser & Cross, 2011). Kaiser and Cross (2011), using the Beginning Teacher Longitudinal Study (BTLS), found that among the beginning public school teachers who were assigned a mentor in 2007–08 about 8 percent were not teaching in 2008–09 and 10 percent were not teaching in 2009–10. In contrast, among the beginning public school teachers who were not assigned a mentor in 2007–08, about 16 percent were not teaching in 2008–09 and 23 percent were not teaching in 2009–10. In addition, in a study of a California school district Villar and Strong (2007) found that the cost associated with mentoring in terms of teacher effectiveness yielded greater savings than the costs associated with attrition.

Some research suggests that significant and demonstrated effects may be more likely when teachers are offered two years of induction support instead of one (Glazerman et al., 2010). A critical review of 15 empirical studies on the topic of induction by the University of Pennsylvania showed positive outcomes in three key areas: 1. job satisfaction, commitment, and retention, 2. classroom instructional practices, and 3. student achievement (Ingersoll & Strong, 2011). Moreover, a federally funded study found that classrooms led by new teachers who were recipients of comprehensive induction support for two years achieved greater student learning gains in mathematics and reading compared to those of new teachers who received less intensive support (Glazerman et al., 2010).

Although still emerging, the research literature emphasizes the importance of obtaining clarity on the goals and purpose of the induction programs (e.g., Arends & Ragazio-Digilio, 2000; Feiman-Nemser, 2001b) and that comprehensive multiyear induction programs have a stronger impact. A report from the New Teacher Center, *Support From the Start: A 50-State Review of Policies on New Educator Induction and Mentoring*, suggested that states that implement strong policy provide the largest amount of support for new teachers (Goldrick, 2016).

A critical feature of induction is the quality of the mentor. Research suggests that states and districts should set criteria for mentor selection and assignment, including the following: interpersonal skills, instructional effectiveness, leadership, work experience, and content-area

and grade-level expertise similar to the mentee's assignment (Alliance for Excellent Education, 2004; Goldrick, Osta, Barlin, & Burn, 2012; Hobson, Ashby, Malderez, & Tomlinson, 2009; Ingersoll & Strong, 2011; Johnson, 2009; Wechsler, Caspary, Humphrey, & Matsko, 2010).

Induction and mentoring can be mechanisms to address the increasing issues of teacher shortages, particularly with special educators and English language learners (Billingsley, Carlson, & Klein, 2004; Brownell, Hirsch, & Seo, 2004; Gandara et al., 2005). Although a considerable induction knowledge base exists in general education, research on induction of special educators is relatively scant. However, there is some emerging evidence that mentoring and induction support influence beginning special educators' intent to remain in teaching (Whitaker, 2000) and perceived effectiveness. Recently, induction has also been linked to beginning teachers' self-ratings of their preparedness to teach, pedagogical content knowledge, and ability to manage classrooms (Boe, Cook, & Sunderland, 2008). In smaller scale studies, beginning teachers' participation in formal induction programs matched to their unique needs influenced their intent to stay in their current teaching positions (Gehrke & McCoy, 2007; Griffin, 2005; Irinaga-Bistolas et al., 2007; Martinez & Mulhall, 2007; Nielsen, Barry, & Addison, 2006; Tucker, 2000).

Improved Partnerships

There is some research to support the strategy of strengthening partnerships between districts and educator preparation programs. A 2014 report from the National Network of State Teachers of the Year and the Center on Great Teachers and Leaders included the results of a survey of national and state teachers of the year about the professional experiences and supports they believed made the greatest contributions to their growth as educators. The survey found that the most important support offered during the preservice stage of teacher development was access to a high-quality clinical practicum (Behrstock-Sherratt, Bassett, Olson, & Jacques, 2014).

As noted earlier, survey data about residencies suggest that a strong partnership between a preparation program and a district leads to increased retention of teachers and employer satisfaction. A report by the Council for the Accreditation of Teacher Education (2010) demonstrated that almost all the Chicago schools that partnered with a residency had increases in the percentage of students meeting or exceeding state standards on achievement. The report included numerous examples of partnerships working to improve school climate, teaching quality, and student achievement.

Policy makers seem to concur that close partnerships matter. The proposed federal regulations that are part of Title II of the Higher Education Act (HEA) related to reporting data on teacher preparation require providers to report data showing that they offer quality clinical preparation and to collect data on employer perceptions of graduates). Some states, such as Massachusetts, require providers to show evidence of partnerships with school districts in order to gain program approval (Massachusetts Department of Elementary and Secondary Education, 2015a). The Council for the Accreditation of Educator Preparation (CAEP), the new accreditor of teacher preparation providers, asks providers for evidence of partnership with school districts in order to earn accreditation (CAEP, n.d.).

Leveraging Educator Evaluation Systems

Despite calls for linking educator evaluation systems to professional development and growth (e.g., Coggshall, Rasmussen, Colton, Milton, & Jacques, 2012; Curtis & Weiner, 2012; Goe, Biggers, & Croft, 2012), the extent to which these kinds of linkages occur in practice is not clear. Most current research focuses on content-based or other types of professional development or feedback and not on specific uses of evaluation data.

Research supports that induction and mentoring can have positive effects on teacher retention and improvements in practice; however, success is largely dependent on the quality of the induction and mentoring programs. While using educator evaluation data to guide professional learning and support provided through induction and mentoring is logical, research supporting this recommendation is not yet available.

Recommendation 2. The Department will define effective teachers as those who strive to engage all students in learning, demonstrate instructional and subject matter competence, and continuously grow and improve.

Under ESSA, requirements related to "highly qualified teachers" have been eliminated, meaning states are free to use any standards for teacher certification. The work group, therefore, felt it important that teacher certification standards in the state reflect a well-rounded definition of an effective educator that can be assessed by appropriate metrics and supported through relevant strategies. The work group discussed a number of approaches that would define an effective educator and determine both metrics and strategies to monitor and promote equitable access to effective educators for all students. These approaches are discussed as sub-recommendations 3a and 3b, as follows:

- **Sub-Recommendation 2a:** Define components of an effective educator.
- **Sub-Recommendation 2b**: Identify metrics and strategies to promote equitable access to effective educators.

Defining Effective Educators

Though significant research has been done on measuring effective teaching, definitions of effective teaching or an effective educator center on theory and beliefs about what makes for successful teaching. Pennsylvania's 2015 Equity Plan simply defines "effective" educators as those whose overall effectiveness rating is "proficient" or "distinguished."

Many states and professional organizations have created their own definitions, which generally speak to multiple elements, such as teachers' contributions to student learning and other student outcomes, their contributions to their profession, knowledge of teaching, and possibly also relationship with parents or the community. For example, the U.S. Department of Education defines effective teaching as: "teachers know how to use the curriculum and instructional assessments to support each child's learning and how to engage families as partners in children's development and education" (U.S. Department of Education, 2012). A report from the Commission on Effective Teachers and Teaching (2012) also places a strong emphasis on student learning, stating that "the ultimate measure of effectiveness is evidence of a teacher's contributions

to student learning and well-being, to the educational community, and to the profession." A research report from National Comprehensive Center for Teacher Quality (Goe, Bell, & Little, 2008) offers the following five factors that states and policymakers should consider when defining effective educators:

- Effective teachers have high expectations for all students and help students learn, as measured by value-added or other test-based measures, alternative or otherwise;
- Effective teachers contribute to positive academic, attitudinal, and social outcomes for students;
- Effective teachers use diverse resources to plan and structure engaging learning opportunities;
- Effective teachers contribute to the development of classrooms and schools that value diversity and civic-mindedness; and
- Effective teachers collaborate with other teachers, administrators, parents, and education professionals to ensure student success.

Metrics and Strategies to Promote Equitable Access to Effective Educators

Once effective teaching is defined, the challenge then becomes how to measure effectiveness appropriately and how to ensure that all students have access to effective educators. Over the last several years, states and many other organizations have invested significant resources into measuring effective teaching (see sections on educator evaluation later in this report for more detail on measures).

States have also committed to ensuring equitable access to effective teaching (as required by the federal government). However, research continues to show that teacher quality gaps are pervasive and more adversely affect low-income schools (Goldhaber, Lavery, & Theobald, 2015). A summary of state equity plans suggest the following leading factors of inequitable access as self-identified by states:

- Novice teachers teach low-performing students at greater rates than experienced teachers.
- When teachers transfer, they generally move into schools that serve students from higher income families than the students they leave behind.
- Teachers working in schools with more disadvantaged students are more likely than other teachers to leave their school districts or transfer.

Evidence exists for some strategies to address these factors, such as improving working conditions (Wynn, Carboni, & Patall, 2007), leadership quality (Leithwood, Seashore Louis, Anderson, & Wahlstrom, 2004; Keithwood, 2004), and compensation factors (Ingersoll, 2003; Keigher & Cross, 2010; Borman & Dowling, 2008), but for others remains sparse.

Recommendation 3. The Department should promote and support collaborative in-field, practical experiences as a crucial component of educator preparation.

In the area of clinical training, the work group discussed the value of clinical experience in preparing and retaining high-quality educators. Members articulated the need to elevate the importance and expectations of clinical experiences across the state within traditional and alternative preparation programs. Most of the work group discussion coalesced around the critical role of partnerships between districts and educator preparation programs, including the role that PDE can play in advancing those expectations. Members raised concern over the lack of quality clinical experiences but noted that extending the time of clinical practice would not necessarily ensure high quality. The work group therefore suggested that program revisions should not extend program requirements but, rather, work to incorporate innovative strategies to integrate deliberate opportunities to practice within existing coursework and field experiences such that program extension does not emerge as a barrier to recruitment.

Pennsylvania statute requires "a well-planned sequence of professional educator courses and field experiences to develop an understanding of the structure, skills, core concepts, facts, methods of inquiry and application of technology related to each academic discipline the candidates plan to teach or in the academic disciplines related to the noninstructional certificate categories in which they plan to serve."(§ 354.25).

In addition, Pennsylvania requires a minimum of 12 weeks of student teaching/clinical training, which requires the completion of four stages of field experience including:

- Observation (the candidate observes a range of education-related settings);
- Exploration (the candidate works under a teacher's guidance with a group of students);
- Pre-student Teaching (the candidate works with groups of students, in school or in an after school settings under the observation of a certified teacher); and
- Teaching (a minimum of 12 weeks full-time student teaching is required).

To this effect, the Educator preparation work group explored and discussed various strategies to improve the quality of their current system, including two distinct sub-recommendations to prioritize within the recommendation.

- **Sub-Recommendation 3a:** Explore and promote models that provide for additional in-field practical experience and real-life situations through the teacher preparation continuum.
- **Sub-Recommendation 3b:** Increase access to quality clinical preparation.

Sub-recommendations are described in more detail below, followed by a summary of relevant research and/or examples of similar current policy or practice.

Sub-Recommendation 3a. *Explore and promote models that provide for additional in-field practical experience and real-life situations through the teacher preparation continuum.*

As is described earlier, currently PDE identifies four stages of field experiences for teaching candidates, including observation, exploration, pre-student teaching, and student teaching. The work group discussed the following strategies in particular to support additional in-field practical experiences and real-life situations for teacher candidates:

- Investigate alternative models concerning the timing and length of clinical experiences; and
- Consider clinical experiences across multiple contexts and settings.

Professional development schools (PDSs) have a long history, reaching back to the ideas of John Dewey and teaching laboratories or lab schools (Dewey, 1896). There is some evidence that teachers prepared in these schools received higher ratings on their practice and tended to view teaching as a career rather than a job (Clift & Brady, 2005). In a PDS model, teacher candidates take all their classes and complete all fieldwork within a single school district.

Aspects of teacher residency programs are somewhat similar in design of PDSs in that they emphasize and provide opportunities to practice within the preparation curriculum. These programs place teacher preparation activities in K–12 schools and bring together mentor teachers and teacher educators in efforts to prepare candidates (Urban Teacher Residency United, 2015). A number of residencies are headquartered outside colleges and universities in not-for-profit organizations or school districts; these often partner with institutions of higher education (IHEs) to provide some of the preparation program. Other residencies are run by IHEs in partnership with a school district.

Recent research indicates that the setting where a teacher completes clinical training has an influence on a variety of teacher employment decisions. One study showed that the setting in which student teaching is completed appears to have an impact on where graduates are hired; graduates are more likely to be hired where they completed student teaching or in a school similar to the one where they completed student teaching (Goldhaber, Krieg, & Theobald, 2013). Another study using data from New York City found that teacher candidates who completed student teaching in schools with lower teacher turnover, also called “easier to staff” schools, were more likely to stay employed for five years (Ronfeldt, 2012). This study also found that student achievement in classrooms taught by teachers who did their clinical training in easier-to-staff schools was higher than in the classrooms of similar teachers who did not. Descriptive and survey data collected by the National Center for Teacher Residencies (NCTR) also suggested that teachers who train in specific urban districts tend to stay there, with 84% of residents still teaching in the districts where they trained (Urban Teacher Residency United [UTRU], 2014). NCTR also surveys hiring principals about its graduates’ effectiveness; 88% said graduates of residencies were more effective than the typical new teacher in instruction and pedagogy and in culturally responsive teaching (UTRU, 2013). Other research suggests that student teachers that did their clinical training in schools with greater teacher collaboration, and to some extent greater teacher retention, were more effective at raising student achievement (Ronfeldt, 2015).

A report from the Collaboration for Effective Educator Development, Accountability and Reform (CEEDAR) Center and the Center on Great Teachers and Leaders (GTL Center) outlined

research-based features for providing high-quality, structured, and sequenced opportunities to practice within teacher preparation programs, including modeling expert teaching, varying settings for clinical experiences, providing coaching and feedback, and having candidates reflect on their teaching (Benedict, Holdheide, Brownell, & Foley, 2016).

Examples of new or relevant approaches used by traditional and alternative programs related to field experiences include the following:

- **The University of Michigan School of Education** has partnered with nearby K-8 schools to create a kind of learning laboratory. Although some teacher interns spend an entire year at the schools in a residency-like program, others work in the school several times a semester throughout their program (DeMonte, 2016). Often teacher interns and university faculty work on a course together in one of the classrooms the schools have devoted to that use. Then they walk down the hall to a classroom and work on the teaching skills they just learned. University faculty and classroom teachers work together closely to create tightly designed teaching experiences that can benefit teacher interns as well as the children in the school.
- In **Arizona State University's iTeachAZ** program, college seniors preparing to be teachers spend the entire school year in a school, working alongside a mentor teacher, rather than completing the more common single semester of student teaching. University faculty teach classes at the school where the candidates are working. Some of the faculty who worked in the iTeachAZ program reported that principals were especially eager to hire new teachers who trained in this program—more so than teachers in more traditional programs (M. Rojas, personal communication, November 30, 2015).
- **St. Cloud University's Academy for Co-Teaching and Collaboration** has created a co-teaching model that has been implemented and studied. In this model, co-teaching is defined as two teachers working together with groups of students and sharing the planning, organization, delivery, and assessment of instruction and the physical space.[27] Rather than one teacher in the lead and the other in a supporting role, two professionals are sharing equally in teaching students in a classroom.

In summary, research supports the important role of field experiences in influencing teacher staffing patterns and, in some cases, effectiveness. There are several innovative programs at IHEs throughout the country that also could be considered when designing robust field experiences for teacher candidates.

Sub-Recommendation 3b: Increase access to quality clinical preparation

The work group discussed the following strategies in particular:

- Strengthen partnerships between districts and educator preparation programs;
- Strengthen training, expectations, and incentives for cooperating teachers; and
- Enhance quality of practice-based opportunities, including opportunities for deliberate and reflective practice.

[27] See http://www.stcloudstate.edu/soe/coteaching/default.aspx

Partnerships

Research related to the strategy of strengthening partnerships between districts and educator preparation programs is discussed in an earlier section.

Mentor or Cooperating Teachers

Research shows the importance of cooperating teachers during student teaching. In one study, Ronfeldt and colleagues found that teacher candidates who reported higher quality mentor teachers and more autonomy to make instructional decisions during their clinical training felt better prepared to teach (Ronfeldt, Reininger, & Kwok, 2013). In addition, the study found that teacher candidates who reported higher quality mentor teachers planned to spend significantly more years teaching in their current district, while teacher candidates who reported more instructional autonomy planned to spend significantly more years in teaching. In another study, Ronfeldt & Reininger (2012) found that student teacher satisfaction and a higher rating on the quality of the cooperating teacher was correlated with a candidate's feeling of preparedness to teach.

Respondents surveyed for the 2014 report from the National Network of State Teachers of the Year and the Center on Great Teachers and Leaders—mentioned earlier—indicated that having a strong cooperating teacher, as defined by the teacher's effectiveness in promoting student learning and providing adult mentorship, was a critical component of a high-quality clinical practicum experience. Survey respondents identified key selection criteria for cooperating teachers, including whether they had received training for the cooperating teacher role, had more than five years of teaching experience, and taught in the same subject area as the student teacher (Behrstock-Sherratt et al., 2014).

Quality of Clinical Training

One study found that the duration of student teaching was not related to teacher candidates' perceptions of preparedness to teach, but that the quality of the clinical training—regardless of duration—had an impact on whether candidates felt ready to teach (Ronfeldt & Reininger, 2012). Another study of national data found that the duration of clinical training was particularly important for teachers with minimal or no coursework on pedagogy (Ronfeldt, Schwartz, & Jacob, 2014). For teachers who had no student teaching, the amount of pedagogical coursework mattered more. This pattern was true for teachers regardless of the selectivity of their program.

Ingersoll, Merrill, and May (2014) conducted a study to determine if the type and amount of education and preparation candidates receive matter. Not surprisingly, the study demonstrates that there is a wide variance in the type of learning experiences teaching candidates receive. That said, the analysis suggest that candidates that receive training in "teaching methods and pedagogy—especially practice teaching, observation of other classroom teaching and feedback on their own teaching—were far less likely to leave teaching after their first year on the job" (Ingersoll, Merril, & May, 2014, p. 1). These findings further support the benefit of clinical practice within preparation coursework and field experiences.

Teacher preparation programs often include coursework along with clinical training, and some research has looked at the relationship between these. One study found that the amount or type of coursework had no effect on a new teacher's ability to increase student achievement (Constantine et al., 2009). A more recent study investigating the perceptions of teacher candidates found that the

link between coursework and clinical experiences contributed to a teacher candidate's sense of how to apply coursework to clinical training (Grossman, Ronfeldt, & Cohen, 2012).

In summary, some research and policies support strong partnerships between IHEs and districts to improve teacher candidate quality. Research shows that having a strong mentor or cooperating teacher can positively impact a teacher candidate, though specific strategies how to improve the training, expectations and incentives for cooperating teachers is still emerging. Finally, there is research on the importance of quality clinical training experiences, but there is less research on what those programs should look like and no research on the role of reflection in clinical experiences.

Recommendation 4. The Department should promote and increase opportunities to recruit, retain, and support diverse and talented school leaders.

Research has shown that school leadership is second only to classroom instruction among factors that contribute to student learning at school. Moreover, the effects of good leadership are largest in the neediest schools (Leithwood et al., 2004). The Wallace Foundation report analyzed the research on leadership and through this analysis suggested three broad categories of successful leaders; 1) setting direction; 2) developing people; and 3) redesigning the organization (Leithwood et al., 2004). Recognizing the essential role of the leader, ESSA includes funding sources that can pay for a state's school leadership improvement efforts.

The work group discussed how to recruit, retain, and support diverse and talented school leaders and made the following four sub-recommendations:

- **Sub-Recommendation 4a.** Create statewide marketing campaign to target audiences; consider incentives.
- **Sub-Recommendation 4b.** Create robust mentoring and coaching programs pre and post hire.
- **Sub-Recommendation 4c.** Use the principal evaluation tool to elevate the importance of ongoing support.
- **Sub-Recommendation 4d.** Provide more pathways within districts for teacher leaders and other educators professional to develop as school leaders.

Statewide Marketing Campaigns

Similar to the situation with teacher marketing campaigns, research on statewide marketing campaigns to recruit effective school leaders is currently not available. Furthermore, recruitment practices are usually the function of local school districts. However, states can play a proactive role in the principal recruitment process through changing the incentives for prospective principals. A report from the Wallace Foundation (Mana, 2015) recommends that states consider the following actions to improve principal recruitment:

- Facilitate coordination between local school districts and principal preparation programs.
- Alter incentives to increase the chances that people who seek principal certification actually intend to become principals.

- Support special institutes, including leadership academies, to help identify potentially talented principals.
- Forecast future trends in anticipated principal vacancies to direct recruitment toward meeting specific state needs for principals.

A program called New Leaders for New Schools has developed marketing and recruitment strategies to attract candidates outside the school system who would make effective school leaders. The program targets those who have led community organizations, nonprofits, and youth development programs. Through creating local, regional, and national networks, half of the program enrollees are from outside public school systems. New Leaders' techniques can be seen as one example of how to expand the talent pool for prospective school leaders (U.S. Department of Education, 2004). In its first eight years, 95% of participants in the New Leaders program went on to hold leadership positions in urban schools (U.S. Department of Education, 2004). A study conducted by the RAND Corporation suggests that principals from the New Leaders program have demonstrated marked improvement in student performance (Burkhauser, Gates, Hamilton, & Ikemoto, 2012).

Robust Mentoring and Coaching Programs

Although significant research may support the claim that effective principals are critical, there is limited evidence about how best to support principals with coaching or mentoring. Research suggests that a core set of principal leadership practices, ranging from human capital management to agenda setting to coaching and instructional leadership, are associated with improved student outcomes (Hallinger and Heck 1998; Knapp et al., 2006; Harris et al., 2010). However, research also indicates that few principals actually engage in these effective practices, spending minimal amounts of time on instructional leadership activities, coaching, and teacher evaluation (Murphy, 1990; Horng et al., 2009; May & Supovitz, 2011; May et al., 2012; Grissom et al., 2013). Furthermore, few principals are trained to coach teachers on instructional improvement, adequately evaluate their progress, or use teacher effectiveness data for human capital decisions (Murphy et al., 2013; Goldring et al., 2014). Therefore, coaching or mentoring focusing on these elements may be warranted.

There is limited research indicating what level of professional development intensity is required to bring about meaningful changes in principal effectiveness. One of the few rigorous evaluations of a principal professional development program with some evidence of effectiveness (Nunnery et al., 2010; Nunnery et al., 2011) was a program that offered approximately 250 hours of formal group training, coaching, and online content over about an 18-month period, suggesting that intensity may be important to success.

One example of a state that developed a strategic plan for improving school leadership is Illinois, which has designed tightly integrated courses that include fieldwork and internship experiences. Their strategic plan underlines the importance of clinical experiences in preparing prospective principals to become effective leaders, stating that site-based learning is one of the most critical components in guaranteeing the success of developing leadership skills within prospective principals. Illinois's strategic plan recommends that the state build its capacity within regions and

districts to develop mentors and coaches who are able to impart their knowledge of leadership to the incoming generation of school leaders (The Wallace Foundation, 2016).

Principal Evaluation Tools and Ongoing Support

Though some research has been done on measuring effective leadership, specific research on how to use principal evaluation information to guide ongoing support is limited to the information described previously about principal professional development in general.

Pathways into School Leadership for Teacher Leaders

Though not much research exists in terms of creating effective pathways within districts for teacher leaders, there are a number of state examples of these types of programs. For instance, the Rhode Island Center for School Leadership focuses on developing and promoting school leaders from within districts. In the program, teachers who have shown leadership potential are identified by their superintendents. After the selection process, the teachers attend a series of workshops, seminars, and panel discussions led by experienced principals in order to provide an overview of school leadership roles. In addition, the selected teachers spend one week working beside a mentor principal at a school of their choice (Education Alliance at Brown University, n.d.).

A report from the U.S. Department of Education looked at six programs that have implemented innovative approaches to recruiting and preparing principals. One noteworthy example is Cleveland's First Ring Leadership Program. Much like the Rhode Island Center for School Leadership, participants in Cleveland's program are selected by superintendents based on their prospective talent to be effective school leaders in the capacity of principal, assistant principal, or designated teacher-leader. The report also highlights New Jersey's EXCEL program, which aims to increase, diversify, and improve the pool of effective school leader candidates. One way the program achieves this is through using superintendents to identify teachers in their districts and support their enrollment into the program with funding and time to participate in coursework and internships (U.S. Department of Education, 2004).

In summary, there are some policy recommendations on principal recruitment and there is one successful program that could be considered an exemplar. There is little research on effective principal professional development, coaching and mentoring practices. Finally, research does not support or negate the recommendation around effective pathways within districts for teacher leaders, however there are some state and local-level examples.

Educator Evaluation

Summary of Findings

In the area of educator evaluation, the work group made two recommendations, the second of which included a broad range of suggestions that we capture in four sub-recommendations. Although the research base pertaining to these recommendations is quite limited and does not offer firm guidance, there is research to support several of the sub-recommendations.

ESSA Requirements

ESSA includes no specific requirements regarding teacher and principal evaluation systems, permitting states and districts to choose how to evaluate teachers and principals. It states: "Nothing in this title shall be construed to authorize the Secretary or any other officer or employee of the Federal Government to mandate, direct, or control a State, local education agency, or school's…teacher, principal, or other school leader evaluation system" (ESSA, 2302(a)(2)).

Even though certain components of educator evaluation systems are no longer required by federal law, educator evaluation remains an important component of a comprehensive human capital management system. Educator evaluation provides a mechanism to assess teacher and principal effectiveness with the goal of supporting their growth and ensuring that students have access to teachers and principals with the skills, knowledge, and dispositions to help them achieve their potential.

Work group members noted a number of concerns with the current educator evaluation system that keep it from being as useful as it could be: its complexity, its limited ability to provide useful feedback due to timeline and training issues; and its lack of applicability to specialized instructional and noninstructional educators. The work group discussed a number of ideas to address these challenges, including streamlining the evaluation process, ensuring the accuracy of evaluation ratings, reconsidering scoring weights, developing tools for specific educator groups, emphasizing feedback that promotes growth through training, concomitant with the evaluation timeline.

Ultimately, the Educator Evaluation work group made the following recommendations:

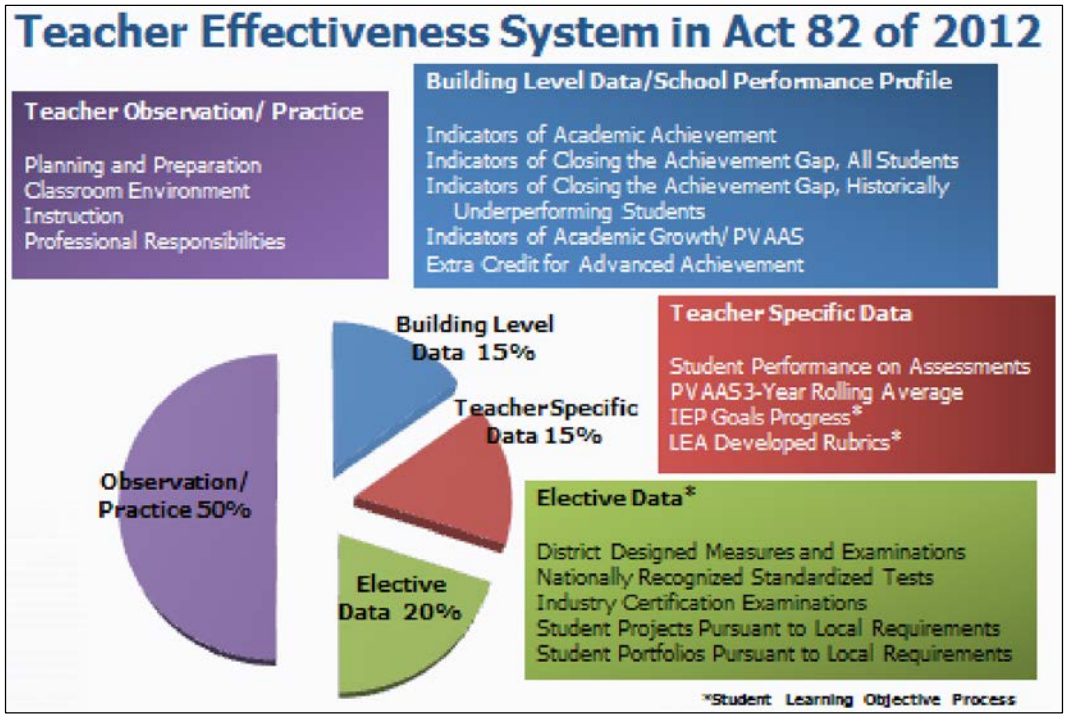
- **Recommendation 1.** Revise the overall components of the professional evaluation systems to reflect the following provisions that support teacher quality and student achievement: 80% professional practice (observation) and 20% student measures (revised SPP or combination of SPP and other relevant data as identified in the LEA's comprehensive plan).
- **Recommendation 2.** Ensure that LEAs implement Pennsylvania's educator evaluation system using a differentiated and collaborative process which promotes educator growth.

More detail on the work group conversations can be found in the notes from work group meetings available at the PDE website (see *Appendix B* for details). In the following sections, we briefly summarize the work group's discussion around the recommendations and the associated research evidence.

Recommendation 1: Revise the overall components of the professional evaluation systems to reflect the following provisions that support teacher quality and student achievement: 80% professional practice (observation) and 20% student measures (revised SPP or combination of SPP and other relevant data as identified in the LEA's comprehensive plan).

The work group discussed perceived challenges with the current educator evaluation system under Act 82—namely, that it contains too many components and gives too much weight to student growth and achievement measures that may lie outside an educator's immediate control. The current teacher evaluation system includes four categories, each containing a number of associated measures (see Figure 3).

Figure 3. Current Teacher Evaluation System Measures (2014)



Source: Pennsylvania Department of Education (2014)

The principal evaluation system is equally complex. Stakeholders noted that the combination of those components and the associated measures is time-consuming for educators to implement and often confusing to understand, thus taking time and resources away from using the system to promote educator growth and development.

In order to ensure a fair, reliable evaluation system that is responsive to local context, the work group recommended reducing the number of categories from four to two and providing school districts with more discretion in how measures are included in the evaluation system. In doing so, the work group noted that districts could choose to include some teacher-specific student data, including student learning objectives, PVAAS rolling averages, and IEP achievement data

in the professional practice process, as a part of pre-observation meetings, post-observation meetings, or summative meetings.

Relevant Research

There is very little research either to support or negate this first recommendation on the number and weights of the various components of the evaluation system. Although the research generally recommends multiple evaluation measures (Little, Goe, & Bell, 2009), it tends not to recommend a specific number or appropriate weights within teacher evaluation measures. Rather, this decision is often seen as one that ought to be based on the values of the individual state or district and their stakeholders (Little, Goe, & Bell, 2009).

One example of research that does recommend evaluation measure weights is the Measures of Effective Teaching (MET) project funded by the Bill & Melinda Gates Foundation (Gates Foundation). The MET project looked at different measures of teacher effectiveness and how to combine them to increase reliability and validity for capturing teacher performance in several school districts, including Charlotte-Mecklenburg (NC), Dallas, Denver, Hillsborough County (FL), New York City, Memphis, and Pittsburgh. The project looked at how well different measures of teacher effectiveness related to student performance; it did *not* study any intervention (i.e., to assess the effect of an evaluation system on teacher performance) but, rather, simply identified correlations across different measures of teacher effectiveness. It is also worth noting that the project allowed remote expert observers to code videos of teacher practice and, therefore, may not be representative of evaluation systems that rely on in-person classroom observations followed by feedback sessions. These caveats notwithstanding, the MET project found that "combining observation scores with evidence of student achievement gains and student feedback" through the use of the TRIPOD student survey "improved predictive power and reliability" of the evaluation system (Kane & Staiger, 2012, p. 9). That means that using these three measures together improves the ability of the evaluation system to correctly identify which teachers will show improved student achievement gains with other students. On the basis of this study, the Gates Foundation recommends the following:

> When combining measures into a single index, we have found that approaches that allocate between 33 percent and 50 percent of the weight to student achievement measures are sufficient to indicate meaningful differences among teachers. Moreover, the authors suggest that balanced weights avoid the risks posed by too narrow a focus on one measure. Overweighting any single measure (whether student achievement, observations, or surveys) invites manipulation and detracts attention and effort away from improvement on the other measures. (2013, pp. 3, 6)

In other words, the combination of measures helps to address any measurement issues inherent in any individual measure. Steinberg and Garrett (2016) similarly noted that there are issues with relying heavily on observation measures to reliably and validly evaluate teacher effectiveness; specifically, they find that incoming student achievement affects a teacher's observation scores.

In spite of the fact there is not strong research indicating how a state or district should weight evaluation measures, research does show that they are important. Analyzing data from the MET project, Steinberg and Kraft (2016) found that "both the weighting schemes assigned to

performance measures and the rating thresholds set by evaluation systems play a critical role in determining teacher proficiency rates" (p. 22).

Given that the MET project represents a single set of data and did not examine an evaluation system intervention, it may have limited relevance for Pennsylvania stakeholders. But it may be useful to consider the data on Pennsylvania's current evaluation system in light of the MET study's recommendation. A study of the piloting of the *Framework for Teaching* in Pennsylvania finds that over 97% of teachers received ratings of *proficient* or *distinguished* on most components and domains on the *Framework for Teaching* (Lipscomb, Terziev & Chaplin, 2015). The study further shows a relationship between professional practice measures and student achievement as measured by value-added scores. Specifically, the researchers found that "teachers with higher component, domain, and [professional practice rating] PPR scores were more likely to have higher value-added scores" (Lipscomb, Terziev & Chaplin, 2015, p. 18). So, though there is not tremendous variation in professional practice scores in Pennsylvania, there is some evidence that these scores are correlated with student growth.

In summary, the Educator Evaluation work group's recommendations to include only two measures and weight the professional practice measures at 80% of an educator's rating may run counter to the best available research. However, given limitations of the research base and stakeholder concerns about the current evaluation system being too unwieldy, such changes might align the educator evaluation system with the values of educators themselves and may ensure that the system is well understood by those educators being evaluated.

Recommendation 2. Ensure that LEAs implement PA's educator evaluation system using a differentiated and collaborative process which promotes educator growth.

The work group expressed concern that the current educator evaluation process did not result in as much growth in teacher and principal practice as it could for a variety of reasons, including lack of position-specific rubrics and challenges with the use of observational rubrics and providing feedback. The work group noted that the educator evaluation system should focus on educator growth and feedback for *all* educators and provide opportunities for collaboration. The work group conversation provided some details for how the educator evaluation system could accomplish this goal, resulting in the following sub-recommendations:

- **Sub-Recommendation 2a:** Include position-specific observation rubrics in the educator evaluation system;
- **Sub-Recommendation 2b:** Rotate educators with no performance concerns through cycles of formal evaluation and supportive growth;
- **Sub-Recommendation 2c:** Assure evaluator competence in the use of observation rubrics; and
- **Sub-Recommendation 2d:** Provide timely, formative feedback.

Sub-recommendations are described in more detail below, followed by a summary of relevant research and/or examples of similar current policy or practice.

Sub-Recommendation 2a: *Include position-specific observation rubrics in the educator evaluation system*

Pennsylvania implemented the *Framework for Teaching* observational system statewide as part of the teacher evaluation system. A key benefit of using the *Framework for Teaching* is that there is a consistent metric for all teachers.

Research on the use of the *Framework for Teaching* is extensive (for example, see Sartain, Stoelinga & Brown, 2009; Taylor & Tyler, 2012; Lipscomb, Terziev & Chaplin, 2015). For example, the MET project found that teachers with higher scores on the *Framework for Teaching* had students with higher value-added scores in English and math (Kane & Staiger, 2012). However, there are no data linking *Framework for Teaching* scores to student achievement outside tested academic subjects. In addition, research by Steinberg and Garrett (2016) analyzing the data of the MET project found that the "incoming achievement of a teacher's students significantly and substantively influences observation-based measures of teacher performance. . . . Specifically, incoming student achievement exerts a larger influence on the measured performance of ELA teachers (compared with math teachers) and subject-matter specialists (compared with their generalist counterparts)" (p. 312).

In contrast, there is no research on position-specific observation rubrics or altered versions of the *Framework for Teaching*. The National Comprehensive Center for Teacher Quality *Practical Guide to Evaluating Teacher Effectiveness* noted that "it is still unclear whether adaptations of the *Framework for Teaching* work as well as the original version" (Little, Goe, & Bell, 2009, p. 6).

In addition to the lack of research supporting the adaptation of the *Framework for Teaching*, there is no research to consult regarding the more general notion that position-specific observation rubrics should be made available. However, several states provide resources for supporting evaluators of teachers in different roles, including the following examples:

- Pennsylvania provides examples and guiding questions for teachers who have unique roles and functions
- Illinois has provided guidance to districts on ensuring that teacher evaluation systems meet the needs of teachers of students with disabilities, teachers of English learners, and teachers of early childhood students. The guidance document does not recommend creating new evaluation rubrics for teachers in these roles (Illinois State Board of Education, 2014).
- The Washington, DC teacher evaluation system (IMPACT) includes rubrics for different roles and responsibilities, including different grade levels and subject specialties. The 26 rubrics from 2015–16 can be found on the IMPACT website.
- Arkansas has provided resources for educators in different roles in a resources handbook.

- Massachusetts has created a Guide to Rubrics and Model Rubrics for Superintendent, Administrator, and Teacher, which includes how districts can choose to modify the state's model rubrics to different roles (Massachusetts Department of Elementary and Secondary Education, 2015b). The Massachusetts Educator Effectiveness Guidebook for Inclusive Practice demonstrates how aspects of the evaluation cycle can reflect inclusive practices.
- The Center on Great Teachers and Leaders created a special issues brief (Holdheide, 2013), which describes how educator evaluation systems can support one subset of teachers—teachers of students with disabilities—and provides examples of how states have chosen to do this.

In summary, although there is no research base to support the use of position-specific rubrics, there are examples of this type of practice in other states.

Sub-Recommendation 2b: *Rotate educators with no performance concerns through cycles of formal evaluation and supportive growth.*

The work group recommends streamlining the evaluation system so teachers and principals with no performance concerns rotate through cycles of evaluation and growth. This approach would reduce the administrative burden on a district by reducing the overall number of observations and evaluations required to be conducted each year, which, in turn, might allow district resources to be targeted to supporting growth.

There is no direct supporting research for this recommendation. Research on the implementation of the *Framework for Teaching* in Cincinnati among mid-career teachers conducted by Taylor and Tyler (2012) lends some support for the notion that annual evaluations for effective educators may not be necessary to achieve growth goals. During the years of the study, tenured teachers were evaluated every five years; yet the authors found that the improvement in practice continued after the year of evaluation. They concluded that "the results of our study provide evidence that subjective evaluation can improve employee performance, even after the evaluation period ends" (Taylor & Tyler, 2012, p. 84).

As with other work group recommendations, the research literature base cannot provide significant guidance. However, other states have also implemented policies similar to what the work group is suggesting, such as the following:

- Ohio currently allows for teacher evaluation every three years for teachers rated *accomplished* and every two years for teachers rated *skilled*.
- Illinois allows for tenured teachers rated excellent or proficient to be evaluated every two years.
- Massachusetts allows educators with a Proficient or Exemplary overall rating to be evaluated every two years, if they have a moderate or high student impact rating (Massachusetts Department of Elementary and Secondary Education, 2011).
- Oklahoma allows career teachers who receive an overall evaluation rating of superior or highly effective to be evaluated once every three years.

Informal information-gathering from these states may provide some guidance in the absence of rigorous research.

Sub-Recommendation 2c: *Assure evaluator competence in the use of observation rubrics.*

The work group spoke to the importance of assuring evaluator competence if the evaluation system is to achieve its objectives and result in growth. On this point, there is research available both to confirm the importance of this recommendation and to provide guidance on how to implement it. The research suggests that assuring evaluator competence is one way to increase teacher confidence in the evaluation system and increase its validity (Archer et al., 2016). As important as valid observation protocols are, "equally important are well-trained and calibrated observers to utilize those instruments in standard ways so that results will be comparable across classrooms" (Little, Goe, & Bell, 2009, 6). Archer et al. (2016) noted, "Recent studies have shown students have greater gains in learning when their teachers receive effective observations and effective feedback" (p. 20; Archer et al., 2016 go on to highlight the studies by Steinberg & Sartain, 2015 and Allen et al., 2011).

There are several research studies that provide recommendations on how to secure and continue ongoing evaluator competence, including the MET project. These studies recommend the following strategies for assuring evaluator competence.

- Conduct initial observer training followed by certification (Little, Goe, & Bell, 2009; Kane & Staiger, 2012; Archer et al., 2016). In terms of how to approach this initial observer training, Cantrell & Kane (2013) noted: "There is great potential in using video for teacher feedback and for the training and assessment of observers" (p. 20). Curtis (2012) describes how Hillsborough County Public Schools conducted initial and ongoing training for evaluators, using an online pre-training, video training, and in-person observations with trainers. Evaluator certification required an in-person observation of evaluators conducting pre-observation conferences, observation, and post-observation conferences (Curtis, 2012).
- Provide ongoing support for evaluators and require periodic recertification (Archer et al., 2016). Ongoing support for evaluators can improve scoring and consistency in observation ratings (Cash, Hamre, Pianta, & Myers, 2012). In addition, initial certification "does not and cannot guarantee that this observer will always demonstrate this level of performance" so a schedule for periodic recertification or retraining should be built into the teacher evaluation system (McClellan, Atkinson, & Danielson, 2012, p. 13).
- Ensure teachers are observed annually by at least one observer outside of the school. In their study of teacher observations in four districts, Whitehurst, Chingos, and Lindquist (2014) noted that the validity of observations is increased when trained observers from outside of the school conduct at least one of a teacher's observations, noting the observer should not have "substantial prior knowledge of the teacher being observed" (Whitehurst et al., 2014, p. 3). The MET study has similar findings: "for the same total number of observations, incorporating additional observers increases reliability" (Gates Foundation, 2013, p. 18).
- Conduct ongoing system monitoring to ensure evaluation and feedback components are being implemented as intended (Archer et al., 2016). For example, districts could do this

by analyzing the scores of teachers in different measures to see if they are generally aligned. Another approach to monitoring is described by the Gates Foundation: "One way to monitor reliability is to have a subset of teachers observed by impartial observers (who come from outside the teachers' school and have no personal relationship to the teachers) and compare the impartial scores with the official scores" (Kane & Staiger, 2012, p. 14). The purpose of these observations in this context would be to monitor the system, not to provide individual teachers with feedback.

In summary, research supports several different approaches to assuring evaluator competence, including initial training, certification, multiple observers and system reliability checks.

Sub-Recommendation 2d: *Provide timely, formative feedback*

The work group also recommended that efforts be made to ensure the evaluation system results in timely, formative feedback. Ensuring that evaluations provide teachers with timely, formative feedback on their practice is one way to increase how well the system supports teachers in improving their practice, and existing research does support this recommendation (see Archer et al., 2016; Allen et al., 2011).

After a review of state teacher evaluation policies, artifact review, interviews, and focus groups, Gandha & Baxter (2016) from the Southern Regional Education Board encourage states to "focus on accurate, practical and timely feedback" as their first recommendation for how to continue to progress in implementing state evaluation systems (p. 6). In an analysis of teacher observations in four districts, the Brown Center on Education Policy at the Brookings Institute highlights the importance of feedback from classroom observations in providing immediate feedback to teachers (Whitehurst et al., 2014). The authors note "Classroom observations have the potential of providing formative feedback to teachers that helps them improve their practice, whereas feedback from state achievement tests is often too delayed and vague to produce improvement in teaching" (Whitehurst et al., 2014, p. 2).

Two studies of feedback through teacher evaluation systems also found evidence of positive effects on student achievement when teachers who were evaluated based on the Framework for Teaching (FFT) observation rubric received structured feedback as part of the district's formal evaluation process. Taylor and Tyler (2012) compared the performance of teachers' students in Cincinnati during the years prior to a teacher's evaluation with their students' performance in the year of the evaluation and the following year; they found an impact on mathematics (but not reading) achievement in the postevaluation years. Steinberg and Sartain (2015) compared student performance in schools before and after they began using the district's new evaluation system and found an impact in reading (and marginally in mathematics) but only for some schools.[28]

More broadly, feedback and its influence on human behavior have been the subject of thousands of studies in multiple disciplines. In a meta-analysis of over 100 studies, Kluger and DeNisi (1996) found that feedback produced negative, rather than positive effects on performance about one-third of the time. To account for this, they hypothesized that feedback changes the locus of attention among task learning, motivation for learning the task, and self-esteem, and that the

effectiveness of feedback decreases as attention moves closer to the self and away from the task. A 2009 review by Brutus takes a more applied perspective to summarize research from industrial–organizational psychology, communication, and human resource management applicable to the use of narrative comments in performance appraisals. Brutus identifies four dimensions of narrative comments—valence, domain coverage, specificity, and the inclusion of suggestions for improvement. Brutus (2009) lays out a series of propositions regarding the ways in which these dimensions will influence the recipient: narrative comments convey more personal focus than ratings and, therefore, will lead to stronger reactions from recipients; the positive influence of narrative units on subsequent behavior will be moderated by their level of specificity such that their influence will be greater when specific; narrative units, in general, will be perceived as less normative and will lead to less social comparison than ratings; comments that include prescriptive information will have a greater impact on behavior change than simple comments that are only descriptive; and the relation of narrative comments on behavior change on a particular domain will be moderated by their variability such that a weaker relationship will be observed when the variability among comments is high.

In a summary of research by the authors and others on the influence of negative and positive feedback on the actions of recipients, Fishbach, Eyal, and Finkelstein (2010) posit that positive feedback functions to increase goal commitment, whereas negative feedback can increase effort by signaling insufficient progress. However, people can be more or less receptive to negative feedback depending on circumstances, including the extent to which they felt that commitment was already assured.

Shute (2009) summarizes research-based guidance for providing formative feedback to learners including focusing the feedback on the task, not the learner, presenting elaborated feedback, but only in manageable units; being specific and clear with feedback messages; reducing uncertainty between performance and goals; and providing feedback after learners have attempted a solution.

In summary, the importance of ensuring that evaluations result in timely feedback for teachers is supported by the existing research base. Indeed, research suggests some evidence about specifics of feedback: the value of keeping feedback focused on the task, not the learner (or self); employing a rubric that can clearly demonstrate the alignment between the teacher's actions and the desired goal (reduce uncertainty between performance and goals); focusing on a few high leverage behaviors so that feedback can be delivered in manageable units; aligning with the district's and school's vision of teaching so that, overall, the teacher does not get conflicting feedback; allowing opportunities for practice between sessions so that feedback can be delivered after the teacher has attempted a solution; and establishing a committed relationship between teacher and coach so that the teacher is more open to processing negative feedback. Translating this evidence very specifically to the Pennsylvania context and successfully implementing such a feedback system, however, may continue to present challenges.

References

Ableidinger, J. (Spring 2015). *A is for affluent.* Public School Forum of North Carolina. Retrieved from http://www.ncforum.org/wp-content/uploads/2013/05/A-is-for-Affluent-Issue-Brief-Format.pdf.

Adams, C., & Forsyth, P. B. (2013). *Oklahoma school grades: Hiding "poor" achievement.* Norman, OK: The Oklahoma Center for Education Policy.

Adamson, F., & Darling-Hammond, L. (2012). Funding disparities and the inequitable distribution of teachers: Evaluating sources and solutions. *Education Policy Analysis Archives 20,* no. 7: 1–46.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary-school instruction and student achievement. *Science, 333*(6045), 1034–1037.

Allensworth, E., Correa, M., & Ponisciak, S. (2008). *From high school to the future: ACT preparation—Too much, too late*. Chicago, IL: Consortium on Chicago School Research at the University of Chicago. Retrieved from https://consortium.uchicago.edu/sites/default/files/publications/ACTReport08.pdf

Alliance for Excellent Education. (2004). *Tapping the potential: Retaining and developing high-quality new teachers.* Washington, DC: Author. Retrieved from http://all4ed.org/reports-factsheets/tapping-the-potential-retaining-and-developing-high-quality-new-teachers/

Americans for the Arts. (2016). *Americans speak out about the arts: An in-depth look at perceptions and attitudes about the arts in America*. New York, NY. Retrieved from http://www.americansforthearts.org/by-program/reports-and-data/research-studies-publications/public-opinion-poll-overview

Amrein, A. L., & Berliner, D. C. (2012). *An analysis of some unintended and negative consequences of high-stakes testing*. Tempe, AZ: Arizona State University Education Policy Studies Laboratory, Education Policy Research Unit (EPRU).

Amrein-Beardsley, A. (2009). The unintended, pernicious consequences of" staying the course" on the United States' No Child Left Behind policy. *International Journal of Education Policy and Leadership*, *4*(6).

Archer, J., Cantrell, S., Holdtzman, S. L., Joe, J. N., Tocci, C. M., & Wood, J. (2016). Better feedback for better teaching: A practical guide to improving classroom observations. Seattle, WA: Bill and Melinda Gates Foundation and Jossey-Bass.

Baker, B., Sciarra, D., & Farroe, D. (2015). *Is school funding fair? A National Report Card*. Newark, NJ: Education Law Center: 28. Retrieved from http://www.schoolfundingfairness.org/National_Report_Card_2015.pdf.

Baker, E., Linn, R., Herman, J., & Koretz, D. (2002). *Standards for educational accountability systems*. Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Testing.

Balfanz, R., & Byrnes, V. (2013). *Meeting the challenge of combating chronic absenteeism: Impact of the NYC Mayor's Interagency Task Force on chronic absenteeism and school attendance and its implications for other cities.* Baltimore, MD: Everyone Graduates Center, Johns Hopkins University School of Education.

Beaudin, B. (1995). Former teachers who return to public schools: District and teacher characteristics of teachers who return to the districts they left, *Educational Evaluation and Policy Analysis 17*, no. 4: 462–475.

Behrstock-Sherratt, E., Bassett, K., Olson, D., & Jacques, C. (2014). *From good to great: Exemplary teachers share perspectives on increasing teacher effectiveness across the career continuum.* Washington, DC: American Institutes for Research.

Benedict, A., Holdheide, L., Brownell, M., & Foley, A. (2016) *Learning to teach practice-based preparation in teacher education.* Washington, DC: American Institutes for Research. Retrieved from http://ceedar.education.ufl.edu/wp-content/uploads/2016/07/Learning_To_Teach.pdf

Berry, B., Daughtrey, A., & Wieder, A. (2010). *Teacher leadership: Leading the way to effective teaching and learning*. Carrboro, NC: Center for Teaching Quality: 1–26, 7.

Billingsley, B. S., Griffin, C. C., Smith, S. J., Kamman, M., & Israel, M. (2009). *A review of teacher induction in special education: Research, practice, and technology solutions*. University of Florida, Gainesville. Retrieved from http://ncipp.education.ufl.edu/files_6/NCIPP_Induc_010310.pdf\

Billingsley, B. S. (2002). *Special education teacher retention and attrition: A critical analysis of the literature.* Gainesville: University of Florida, Center on Personnel Studies in Special Education.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Blanc, S., Christman, J., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education*, *85*(2), 205-225.

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

Booker, K., & Glazerman, S. (2009). *Effects of the Missouri Career Ladder Program on teacher mobility*, Mathematica Policy Research, Inc.: 1–27, 13.

Boser, U. (2014). *Teacher diversity revisited: A state-by-state analysis.* Center for American Progress. Retrieved from https://cdn.americanprogress.org/wp-content/uploads/2014/05/TeacherDiversity.pdf

Boyle, A., Taylor, J., Hurlburt, S., & Soga, K. (2010). *Title III Accountability: Behind the numbers*. Washington, DC: American Institutes for Research. Retrieved from https://www2.ed.gov/rschstat/eval/title-iii/behind-numbers.pdf

Braun, H., Chudowsky, N., & Koenig, J. (Eds.). (2010). *Getting value out of value-added: Report of a Workshop*. Washington, DC: Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council and National Academy of Education, The National Academies Press. Retrieved from https://www.nap.edu/catalog/12820/getting-value-out-of-value-added-report-of-a-workshop

Briggs, D. C. (2001). The effect of admissions test preparation: Evidence from NELS-88. *Chance*, *14*(1), 10-18.

Brown, R.S., Wohlstetter, P., & Liu, S. (2008). Developing an indicator system for schools of choice: A balanced scorecard approach. *Journal of School Choice*, *2*(4), 392-414.

Brutus, S. (2009). Words versus numbers: A theoretical exploration of giving and receiving narrative comments in performance appraisal. *Human Resource Management Review*, *20*, 144–157.

Burkhauser, S., Gates, S., Hamilton, G., & Ikemoto, G. (2012) *First-year principals in urban school districts: How actions and working conditions*. Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/content/dam/rand/pubs/technical_reports/2012/RAND_TR1191.pdf

Burton, J., Horowitz, R., & Abeles, H. (1999). Learning in and through the arts. In E. Fiske (Ed.), *Champions of change: The impact of the arts on learning*. Washington, DC: The Arts Education Partnership and the President's Committee on the Arts and the Humanities. Retrieved from http://artsedge.kennedy-center.org/champions/pdfs/champsreport.pdf.

California Department of Education. (2016). Letter to the U.S. Department of Education. Retrieved from http://www.cde.ca.gov/nr/el/le/yr16ltr0113.asp

Cantrell, S., & Kane, T. J. (2013, January). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/05/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis*, *33*(3), 378–398.

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*(3), 529–542. Retrieved from http://www.sciencedirect.com/science/article/pii/S0885200611000974

Center for Educator Recruitment, Retention, & Advancement. (2011). *The South Carolina teaching fellows program.* Retrieved from http://www.cerra.org/uploads/1/7/6/8/17684955/tfreport.pdf

Chang, H. N., & Romero, M. (2008). *Present, engaged, and accounted for: The critical importance of addressing chronic absence in the early grades. Report*. New York, NY: National Center for Children in Poverty.

Chester, M. D. (2005). Making valid and consistent inferences about school effectiveness from multiple measures. *Educational Measurement: Issues and Practice, 24*(4), 40–52.

Choppin, J. (April 2002). *Data use in practice: Examples from the school level*. Paper presented at the Annual Conference of the American Educational Research Association. New Orleans, LA.

Chudowsky, N., & Chudowsky, V. (2010). *Rising scores on state tests and NAEP*. Washington, DC: Center on Education Policy. Retrieved from http://files.eric.ed.gov/fulltext/ED513962.pdfa

Cimbricz, S. (2002). State-mandated testing and teachers' beliefs and practice. *Education Policy Analysis Archives, 10*(2), 1–21.

Clift, R., & Brady, P. (2005). Research on methods courses and field experiences. In M. Cochran-Smith & K. Zeichner (Eds.), *Studying teacher education: The report on the AERA Panel on Research and Teacher Education* (pp. 309–324). Mahwah, NJ: Erlbaum.

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

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

Coburn, C., Honig, M., & Stein, M. K. (2005). What's the evidence on districts' use of evidence? Chapter prepared for conference volume, sponsored by the MacArthur Network on Teaching and Learning.

Collier, V. P. (1995). Acquiring a second language for school. *Directions in Language & Education, 1*(4).

Commission on Effective Teachers and Teaching. (2012). *Transforming teaching: Connecting professional responsibility with student learning.* Retrieved from http://www.nea.org/assets/docs/Transformingteaching2012.pdf

Conley, D. T., & Darling-Hammond, L. (2013). *Creating systems of assessment for deeper learning*. Stanford, CA: Stanford Center for Opportunity Policy in Education

Constantine, J., Player, D., Silva, T., Hallgren, K., Grider, M., & Deke, J. (2009). *An evaluation of teachers trained through different routes to certification, final report* (NCEE 2009-4043). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/pubs/20094043/pdf/20094043.pdf

Cordray, D., Pion, G., Brandt, C., Molefe, A., & Toby, M. (2012). *The impact of the Measures of Academic Progress (MAP) Program on student reading achievement*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://files.eric.ed.gov/fulltext/ED537982.pdf

CORE Districts. (2016). *CORE's Social Emotional Learning (SEL) pilot effort*. Retrieved from http://coredistricts.org/social-emotional-learning-efforts/

Council for the Accreditation of Educator Preparation. (n.d.). *Standard 2: Clinical partnerships and practice* [Webpage]. Retrieved from http://caepnet.org/standards/standard-2

Council of Chief State School Officers. (2014). Criteria for procuring and evaluating high-quality assessments. Retrieved from http://www.ccsso.org/Documents/2014/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf

Cowart, S. K. (2010). Driving improvement with a balanced scorecard, *The School Administrator, 67*(2), 16–19.

Curtis, R. (2012). *Building it together: The design and implementation of Hillsborough County Public Schools' teacher evaluation system*. The Aspen Institute. Retrieved from http://communication.sdhc.k12.fl.us/eethome/casestudies/Building%20it%20Together_Aspen%20Case%20Study.pdf

Darling-Hammond, L., Chung, R., & Frelow, R. (2002). Variation in teacher preparation: How well do different pathways prepare teachers to teach? *Journal of Teacher Education 53*(4), 286–302.

Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, J. L., Baker, E., et al. (2013). *Criteria for high-quality assessment.* Stanford, CA: Stanford Center for Opportunity Policy in Education. Retrieved from https://edpolicy.stanford.edu/publications/pubs/847

Darling-Hammond, L., & Pecheone, R. (2010, March). *Developing an internationally comparable balanced assessment system that supports high-quality learning. Paper presented at the National Conference on Next-Generation K–12 Assessment Systems.* Retrieved from http://www.ets.org/Media/Research/pdf/Darling-HammondPechoneSystemModel.pdf

Darling-Hammond, L., Wilhoit, G., & Pittenger, L. (2014). Accountability for college and career readiness: Developing a new paradigm. *Education Policy Analysis Archives, 22*(86), 1–34.

Data Quality Campaign. (2016). Time to act: Making data work for students. Washington, DC. Retrieved from http://dataqualitycampaign.org/event/time-act-making-data-work-students/

Dee, T. S. (2004). MIT Press Journals. *Teachers, race, and student achievement in a randomized experiment.* Retrieved from http://faculty.smu.edu/millimet/classes/eco7321/papers/dee01.pdf

Dembosky, J. W., Pane, J. F., Barney, H., & Christina, R. (2005). *Data-driven decision making in southwestern Pennsylvania school districts* (WR-326-HE/GF). Santa Monica, CA: RAND Corporation.

Deming, D., Cohodes, S., Jennings, J., & Jencks, C. (2016). When does accountability work? Education Next. Retrieved from http://educationnext.org/when-does-accountability-work-texas-system/.

DeMonte, J. (2016). Toward better teacher prep. *Educational Leadership, 73*(8), 66–71.

Dewey, J. (1896). The university school. *University Record,* 5, 417–442.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237–251.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child development*, *82*(1), 405–432.

Dymnicki, A., Sambolt, M., & Kidron, Y. (2013, March). *Improving college and career readiness by incorporating social and emotional learning.* Washington, DC: College and Career Readiness and Success Center. Retrieved from http://www.ccrscenter.org/sites/default/files/Improving%20College%20and%20Career%20Readiness%20by%20Incorporating%20Social%20and%20Emotional%20Learning_0.pdf

Education Alliance at Brown University. (n.d.). *Making the case for principal mentoring.* Retrieved from https://www.brown.edu/academics/education-alliance/sites/brown.edu.academics.education-alliance/files/publications/prncpalmntrg.pdf

Education Commission of the States. (2013). School accountability report cards. Retrieved from http://www.ecs.org/state-school-accountability-report-cards-state-profiles

Ehren, M. C. M., & Star, J. (2013, April). Strategies teachers use to coach students on the math state test. Paper presented at the 94th annual meeting of the American Educational Research Association (AERA), San Francisco, CA.

Faria, A. M., Heppen, J., Li, Y., Stachel, S., Jones, W., Sawyer, K., et al. (2012). *Charting success: Data use and student achievement in urban schools*. Council of the Great City Schools. Retrieved from http://files.eric.ed.gov/fulltext/ED536748.pdf

Faxon-Mills, S., Hamilton, L. S., Rudnick, M., & Stecher, B. M. (2013). *New assessments, better instruction? Designing assessment systems to promote instructional improvement.* Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/pubs/research_reports/RR354.html

Feldman, J., & Tung, R. (2001). Using data based inquiry and decision making to improve instruction. *ERS Spectrum, 19*(3), 10–19.

Feng, L. (2014). Teacher placement, mobility, and occupational choices after teaching. *Education Economics, 22*(1), 24–47.

Florida Department of Education. (2016). *Accountability rules: 2015 Rule development*. Retrieved from http:// www.fldoe.org/accountability/accountability-reporting/accountability-rules.stml

Follman, J. (1992). Secondary school students' ratings of teacher effectiveness. *The High School Journal, 75*(3), 168–178.

Follman, J. (1995). Elementary public school pupil rating of teacher effectiveness. *Child Study Journal, 25*(1), 57–78.

Foundation for Florida's Future. (2014). *Florida formula for student achievement; school grades Q&A.* Tallahassee, FL: Author.

Gandha, T., & Baxter, A. (2016). *State actions to advance teacher evaluation*. Atlanta, GA: Southern Regional Education Board. Retrieved from http://publications.sreb.org/2016/160210_stateactionstoadvanceteachereval.pdf.

Gandara, P., Maxwell-Jolly, J., & Driscoll, A. (2005). *Listening to teachers of English language learners: A survey of California teachers' challenges, experiences, and professional development needs.* Santa Cruz, CA: The Center for the Future of Teaching and Learning. Retrieved from http://www. cftl.org/documents/2005/listeningforweb.pdf

Gates Foundation. (2013). *Feedback for better teaching: Nine principles for using measures of effective teaching.* Seattle, WA: author. Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/05/MET_Feedback-for-Better-Teaching_Principles-Paper.pdf.

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., et al. (2010). *Impacts of comprehensive teacher induction: Final results from a randomized controlled study* (NCEE 2010-4027). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/pubs/20104027/pdf/20104027.pdf

Glazerman, S., Protik, A., Teh, B., Bruch, J., & Max, J. (2013). *Transfer incentives for high performing teachers: Final results from a multisite experiment* (NCEE 2014-4003). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://ies.ed.gov/ncee/pubs/20144003/pdf/20144003.pdf

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://www.gtlcenter.org/sites/default/files/docs/EvaluatingTeachEffectiveness.pdf

Goertz, M. E., & Duffy, M. C., (2000). *Assessment and accountability in the 50 states: 1999–2000*. Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education

Goertz, M. E., Nabors Olah, L., & Riggan, M. (2009). *From testing to teaching: The use of interim assessments in classroom instruction*. Philadelphia, PA: Consortium for Policy Research in Education. Retrieved from http://repository.upenn.edu/cgi/viewcontent.cgi?article=1023&context=cpre_researchreports

Goldhaber, D. (2006). *Everyone's doing it. But what does teaching testing tell us about teacher effectiveness.* University of Washington and the Urban Institute. Retrieved from http://public.econ.duke.edu/~staff/wrkshop_papers/2006-07Papers/Goldhaber.pdf

Goldhaber, D., Lavery, L., & Theobald, R. (2015). *Uneven playing field? Assessing the teacher quality gap between advantaged and disadvantaged students.* Retrieved from http://edr.sagepub.com/content/early/2015/06/29/0013189X15592622.full.pdf

Goldhaber, D., Krieg, J., & Theobald, R. (2013). Knocking on the door to the teaching profession: Modeling the entry of prospective teachers into the workforce. *Economics of Education Review, 43,* 106–124.

Goldhaber, D., & Theobald, R. (2013). *Do different value-added models tell us the same things?* Stanford, CA: Carnegie Foundation for the Advancement of Teaching. Retrieved from http://www.carnegieknowledgenetwork.org/briefs/value-added/different-growth-models/

Goldhaber, D., Theobald, R., & Tien, C. (2015). *The theoretical and empirical arguments for diversifying the teacher workforce: A review of the evidence.* Center for Education Data and Research. Retrieved from http://m.cedr.us/papers/working/CEDR%20WP%202015-9.pdf

Goldrick, L. (2016). *Support from the start: A 50-state review of policies on new educator induction and mentoring.* New Teacher Center. Retrieved from https://newteachercenter.org/wp-content/uploads/2016CompleteReportStatePolicies.pdf

Goodnough, A. (1999). Answers allegedly supplied in effort to raise test scores. New York Times. Retrieved from http://partners.nytimes.com/library/national/regional/120899ny-cheat-edu.html

Gorard, S., Siddiqui, N., & See, B. H. (2015). *Philosophy for children: Evaluation report and executive summary*. London: EEF.

Grossman, P., Ronfeldt, M., & Cohen, J. (2012). The power of setting: The role of field experience in learning to teach. In K. Harris, S. Graham, T. Urdan, A. Bus, S. Major, & H. L. Swanson (Eds.), *American Psychological Association (APA) educational psychology handbook: Applications to teaching and learning* (Vol. 3, pp. 311–334). Washington, DC: American Psychological Association.

Hall, D. (2013). A step forward or a step back? State accountability in the waiver era. Washington, DC: The Education Trust. Retrieved from https://edtrust.org/resource/a-step-forward-or-a-step-back-state-accountability-in-the-waiver-era/

Hamilton, L., Stecher, B., Marsh, J., McCombs, J., Robyn, A., Russel, J., et al. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states*. Santa Monica, CA: RAND Corporation. Retrieved from www.rand.org/pubs/monographs/MG589.html

Hannaway, J. (2007, November 17). *Unbounding rationality: Politics and policy in a data rich system* [Mistisfer Lecture, the University Council of Education Administration].

Hanushek, E., & Raymond, M. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2) 297–327.

Hanushek, E., & Raymond, M. (2006). School accountability and student performance. Federal Reserve Bank of St. Louis. *Regional Economic Development, 2*(1), 51–61.

Harrison, C. (2005). Teachers developing assessment for learning: Mapping teacher change. *Teacher Development*, *9*(2), 255–263.

Harris, D. N., Rutledge, S. A., Ingle, W. K., & Thompson, C. C. (2010). Mix and match: What principals really look for when hiring teachers. *Education Finance and Policy, 5*(2), 228–46.

Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., & Spurgeon, L. (2015). *Student testing in America's Great City Schools: An inventory and preliminary analysis*. Washington, DC: Council of the Great City Schools. Retrieved from http://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/87/Testing%20Report.pdf

Hein, V., Smerdon, B., & Sambolt, M. (2013). Predictors of postsecondary success. Washington, DC: College and Career Readiness and Success Center. Retrieved from https://www.cde.state.co.us/postsecondary/americaninstitutesforresearchpredictorsofpostsecondarysuccess

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *Measuring how benchmark assessments affect student achievement*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2007039_sum.pdf

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2008). *A second follow-up year for measuring how benchmark assessments affect student achievement*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://files.eric.ed.gov/fulltext/ED501327.pdf

Heritage, M. (2011). Formative assessment: An enabler of learning. *Better: Evidence-based Education Magazine*, *3*(3), 18–19.

Herman, J. L. (2010). *Coherence: Key to next generation assessment success.* Los Angeles, CA: Assessment and Accountability Comprehensive Center. Retrieved from https://www.cse.ucla.edu/products/policy/coherence_v6.pdf

Herman, J. L., Baker, E. L., & Linn, R. L. (2004). *Accountability systems in support of student learning: Moving to the Next Generation. CRESST Line— Newsletter of the National Center for Research on Evaluation, Standards, and Student Testing*. Los Angeles, Calif.: UCLA Center for the Study of Evaluation

Hill, R., & DePascale, C. (2002). *Determining the reliability of school scores*. Dover, NH: The Center for Assessment.

Holdheide, L. (2013). *Inclusive design: Building educator evaluation systems that support students with disabilities (revised edition)*. Washington, DC: Center on Great Teachers and Leaders. Retrieved from http://www.gtlcenter.org/sites/default/files/GTL_Inclusive_Design.pdf

Howe, K. R., & Murray, K. (2015). Why school report cards merit a failing grade. *Boulder, CO: National Education Policy Center. Retrieved from http://www.greatlakescenter.org/docs/Policy_Briefs/Howe-StateReportCards.pdf*

Huberman, M., Dunn, L, Stapleton, J., & Parrish, T. (2008). Evaluation of Arizona's system of support: Final report. Washington, DC: American Institutes for Research.

Huntley, B., Engelbrecht, J., & Harding, A. (2009). Can multiple choice questions be successfully used as an assessment format in undergraduate mathematics? *Pythagoras, 69*, 3–16.

Illinois State Board of Education. (2014). *Guidance on building teacher evaluation systems for teachers of students with disabilities, English learners and early childhood students*. Springfield, IL: Author. Retrieved from http://www.isbe.net/peac/pdf/guidance/14-3-teacher-eval-sped-ell-preschool.pdf

Indiana Department of Education. (n.d.). *A–F school accountability FAQ*. Indianapolis, IN: Author. Retrieved from http://www.svcs.k12.in.us/Downloads/revised-f-faq-101712.pdf

Ingersoll, R. (2012). *Beginning teacher induction: What the data tell us.* Kappan Magazine. Retrieved from http://www.gse.upenn.edu/pdf/rmi/PDK-RMI-2012.pdf

Ingersoll, R. (2003). *Is there really a teacher shortage?* Seattle, WA: Center for the Study of Teaching and Policy. Philadelphia, PA: Consortium for Policy Research in Education. Retrieved from https://depts.washington.edu/ctpmail/PDFs/Shortage-RI-09-2003.pdf

Ingersoll, R., & May, H. (2016). *Minority teacher recruitment, employment, and retention: 1987 to 2013*. Learning Policy Institute. Retrieved from https://learningpolicyinstitute.org/sites/default/files/product-files/Minority_Teacher_Recruitment_Employment_Retention%20_BRIEF.pdf

Ingersoll, R., Merrill, L., & May, H. (2014). *What are the effects of teacher education and preparation on beginning teacher attrition?* Research Report (#RR-82). Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania.

Ingersoll, R., & Smith, T. M. (2004). *Do teacher induction and mentoring matter?* Retrieved from http://repository.upenn.edu/gse_pubs/134

Ingersoll, R., & Strong, M. (2011). *The impact of induction and mentoring programs for beginning teachers: A critical review of the research.* University of Pennsylvania Department of Education. Retrieved from http://repository.upenn.edu/cgi/viewcontent.cgi?article=1127&context=gse_pubs

Ingram, D., Seashore, K., & Schroeder, R. G. (2004). Accountability policies and teacher decision making: Barriers to the use of data to improve practice. *Teachers College Record, 106*(6), 1258–1287.

Jacob, B. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*(5–6), 761–796.

Jacob, B. A., & Levitt, S. D. (2002). Rotten apples: An investigation of the prevalence and predictors of teacher cheating, *Quarterly Journal of Economics, 118*(3), 843–877.

Jerald, C. D., Doorey, N. A., & Forgione Jr., P. D. (2011, February). Putting the pieces together: Summary report of the invitational research symposium on through-course summative assessments. Paper presented at the invitational research symposium on Through-Course Summative Assessments, Atlanta, GA. Retrieved from https://www.ets.org/Media/Research/pdf/TCSA_Symposium_Final_Summary.pdf

Kaiser, A., & Cross, F. (2011). *Beginning teacher attrition and mobility: Results from the first through third waves of the 2007–08 Beginning Teacher Longitudinal Study* (NCES 2011-318). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures, *Journal of Economic Perspectives, 16*(4), 91–114.

Kane, T. J., & Staiger, D. O. (2012, January). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/12/MET_Gathering_Feedback_Research_Paper.pdf

Katz, I. R., Friedman, D. E., Elliot Bennett, R., & Berger, A. E. (1996). *Differences in strategies used to solve stem-equivalent constructed-response and multiple-choice SAT-mathematics items.* New York, NY: The College Board. Retrieved from https://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-1995-3-strategies-stem-response-multiple-choice-sat-math.pdf

Kennedy, M. M. (1982). Evidence and decision. In M. M. Kennedy (Ed.), *Working knowledge and other essays* (pp. 59–103). Cambridge, MA: The Huron Institute.

Khanna, R., Trousdale, D., Penuel, W. R., & Kell, J. (1999). *Supporting data use among administrators: Results from a data planning mode*l. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec.

Klute, M. M., Welp, L. C., Yanoski, D. C., Mason, K. M., & Reale, M. L. (2016). *State policies for intervening in chronically low-performing schools: A 50-state scan* (REL 2016–131). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Central. Retrieved from http://ies.ed.gov/ncee/edlabs.

Konstantopoulos, S., Li, W., Miller, S. R., & van der Ploeg, A. (2016). Effects of interim assessments across the achievement distribution evidence from an experiment. *Educational and Psychological Measurement, 76*(4), 587–608.

Koretz, D. (2009). *Measuring up–What Education assessment really tells us*. Cambridge, MA: Harvard University Press.

Koretz, D., Mitchell, K., Barron, S., & Keith, S. (1996). *Final report: Perceived effects of the Maryland school performance assessment program*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing at the UCLA Graduate School of Education. Retrieved from http://cresst.org/wp-content/uploads/TECH409.pdf

Ladd, H., & Sorensen, L. (2016). Returns to teacher experience: Student achievement and motivation in middle school, *Education Finance and Policy*.

Lane, S., Parke, C. S., & Stone, C. A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment, 8*(4), 279–315.

Lazarín, M. (2014). *Testing overload in America's schools*. Washington, DC: Center for American Progress. Retrieved from https://cdn.americanprogress.org/wp-content/uploads/2014/10/LazarinOvertestingReport.pdf

Le Floch, K. C., & Tanenbaum, C. (2016). *Implementing accountability and supports under ESEA flexibility*. Washington, DC: U.S. Department of Education, Office of Planning Evaluation and Policy Development, Policy and Program Studies Service.

Le Floch, K. C., Barnes, C., Massell, D., Boyle, A., Therriault, S., Taylor, J., et al. (2011). *Evaluation of Michigan's system of support for high priority schools: Year 1 report*. Washington, DC: American Institutes for Research.

Leithwood, K., Seashore Louis, K., Anderson, S., and Wahlstrom, K. (2004). *Review of research: How leadership influences student learning.* Retrieved from http://www.wallacefoundation.org/knowledge-center/Documents/How-Leadership-Influences-Student-Learning.pdf

LiCalsi, C., & Piriz, D. (2016). *Evaluation of level 4 school turnaround efforts in Massachusetts, Part 2: Impact of school redesign grants.* Washington, DC: American Institutes for Research

Linn, R. L. (1991). Dimensions of thinking: Implications for testing. In B. F. Jones & L. Idol (Eds.), *Educational values and cognitive instruction: Implications for reform* (pp. 179–208). Hillsdale, NJ: Erlbaum.

Linn, R. L. (2005). Issues in the design of accountability systems. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data in accountability testing. Yearbook of the National Society for the Study of Education* (Vol. 104, Part I, pp. 78–98). Boston, MA: Blackwell Publishing.

Linn, R. L. (2006, June). *Educational accountability systems* (CRESST Technical Report 687). Los Angeles, CA: University of California at Los Angeles, Center for Research on Evaluation, Standards, and Student Testing.

Linn, R. L. (2008). Methodological issues in achieving school accountability. *Journal of Curriculum Studies, 40*(6), 699–711.

Lipscomb, S., Terziev, J., & Chaplin, D. (2015). *Measuring teachers' effectiveness: A report from phase 3 of Pennsylvania's pilot of the Framework for Teaching*. Washington, DC: Mathematica. Retrieved from https://www.mathematica-mpr.com/our-publications-and-findings/publications/measuring-teachers-effectiveness-a-report-from-phase-3-of-pennsylvanias-pilot-of-the-framework

Little, O., Goe, L., & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness*. Naperville, IL: National Comprehensive Center for Teacher Quality. Retrieved from http://files.eric.ed.gov/fulltext/ED543776.pdf

Loukas, A. (2007). What is school climate? High-quality school climate is advantageous for all students and may be particularly beneficial for at risk students. *Leadership Compass, 5*(1), 1–3.

Lukhele, R., Thissen, D., & Wainer, H. (1993). *On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests.* Princeton, NJ: Educational Testing Service. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.1993.tb01517.x/abstract

Marsh, E. J., & Cantor, A. D. (2014). Learning from the test: Dos and don'ts for using multiple-choice tests. In M. A. McDaniel, R. F. Frey, S. M. Fitzpatrick, & H. L. Roediger (Eds.), *Integrating cognitive science with innovative teaching in STEM disciplines* (pp. 37–52). Saint Louis, MO: Washington University in St. Louis.

Marshal, B., & Drummond, M. J. (2006). How teachers engage with assessment for learning: Lessons from the classroom. *Research Papers in Education, 21*(2), 133–149.

Martin, C., Sargrad, S., & Batel, S. (2016). *Making the grade: A 50-state analysis of school accountability systems*. Washington, DC: Center for American Progress. Retrieved from https://www.americanprogress.org/issues/education/report/2016/05/19/137444/making-the-grade/

Marzano, R. J., & Toth, M. (2013). *Teacher evaluation that makes a difference.* Alexandria, VA: ASCD.

Massachusetts Department of Elementary and Secondary Education. (2011). *Overview of the new Massachusetts educator evaluation framework.* Retrieved from http://www.doe.mass.edu/edeval/101511Overview.pdf

Massachusetts Department of Elementary and Secondary Education. (2015a). *Educator preparation* [Webpage]. Retrieved from http://www.doe.mass.edu/edprep/pr.html

Massachusetts Department of Elementary and Secondary Education. (2015b). *Part III: Guide to rubrics and model rubrics for superintendent, administrator, and teacher*. Boston, MA: Author. Retrieved from http://www.doe.mass.edu/edeval/model/PartIII.pdf

McCaffrey, D. F., & Hamilton, L. S. (2007). *Value-added assessment in practice: Lessons from the Pennsylvania value-added assessment system pilot project* (Vol. 506). Rand Corporation.

McClellan, C., Atkinson, M., & Danielson, C. (2012). *Teacher evaluator training & certification: Lessons learned from the Measures of Effective Teaching project (Practitioner Series for Teacher Evaluation).* San Francisco, CA: Teachscape.

McCombs, J. S., Kirby, S. N., Barney, H., Darilek, H., & Magee, S. (2005). *Achieving state and national literacy goals: A long uphill road*. Santa Monica, CA: RAND.

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger III, H. L., & McDaniel, M. A. (2014). *Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes*. St. Louis, MO: Washington University in St. Louis. Retrieved from https://pages.wustl.edu/memory/mcdermott-et-al.-2014

McManus, S. (Ed.). (2008). *Attributes of effective formative assessment*. Washington, DC: Council of Chief State School Officers. Retrieved from http://www.ccsso.org/documents/2008/attributes_of_effective_2008.pdf

Measures of Effective Teaching Project. (2012). *Asking students about teaching: Student perception surveys and their implementation* (Policy & Practice Brief). Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Asking_Students_Practitioner_Brief.pdf

Meyers, R. H. (2000). Value-added indicators: A powerful tool for evaluating science and mathematics programs and policies. *NISE Brief, 3*(3). Madison: University of Wisconsin-Madison, National Center for Improving Science Education.

Mikulecky, M., & Christie, K. (2014, May). *Grading schools: What parents and experts say States should consider to make school accountability systems meaningful.* Denver, CO: Education Commission of the States.

Mishkind, A. (2014, September). Overview: State definitions of college and career readiness. Washington, DC: College and Career Readiness and Success Center. Retrieved from http://www.ccrscenter.org/sites/default/files/CCRS%20Defintions%20Brief_REV_1.pdf National Conference of State Legislatures. (2016). *Summary of the Every Student Succeeds Act, legislation reauthorizing the Elementary and Secondary Education Act*. Retrieved from http://www.ncsl.org/documents/educ/ESSA_summary_NCSL.pdf

National Network of Business and Industry Associations. (2014). *Common employability skills: A foundation for success in the workplace*. Retrieved from http://businessroundtable.org/sites/default/files/Common%20Employability_asingle_fm.pdf

Nayar, N. (2015). *How are states reporting on college and career readiness?* Washington DC: American Institutes for Research. Retrieved from http://www.ccrscenter.org/sites/default/files/AskCCRS_Metrics.pdf

New Mexico Public Education Department. (2016). *NM Teach.* Retrieved from http://ped.state.nm.us/ped/NMTeach_EvaluationPlan.html

Northwest Evaluation Association. (2016). *Make assessment work for all students: Multiple measures matter*. Portland, OR: Author. Retrieved from https://www.nwea.org/content/uploads/2016/05/Make_Assessment_Work_for_All_Students_2016.pdf

Organisation for Economic Cooperation and Development. (2014). *TALIS 2013 results: An international perspective on teaching and learning*. Paris, France: Author.

Pennsylvania Department of Education. (2007). *Pennsylvania's statewide system of school support.* Retrieved from http://www.education.pa.gov/Documents/Teachers-Administrators/Federal%20Programs/School%20Improvement/PA%20SSOSS%20Document.pdf

Pennsylvania Department of Education. (2014a). *ESEA Flexibility Waiver, Title I School Identification Designations*. Retrieved from http://www.pps.k12.pa.us/cms/lib07/PA01000449/Centricity/Domain/145/Title%20I%20School%20Identification%20Designations.pdf

Pennsylvania Department of Education. (2014b). Highlights from Pennsylvania's ESEA Flexibility Request. Retrieved from http://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/School%20Performance%20Profile/PA%20Highlights%20-%20ESEA%20Flexibility.pdf

Pennsylvania Department of Education. (2016a). *2016 Pennsylvania system of school assessment, handbook for assessment coordinators*. Retrieved from http://www.education.pa.gov/K-12/Assessment%20and%20Accountability/PSSA/Pages/Test-Administrators-Materials.aspx#tab-1

Pennsylvania Department of Education. (2016b). *Certification testing.* Harrisburg, PA: Author. Retrieved from http://www.education.pa.gov/Teachers%20-%20Administrators/Certifications/Pages/Certification-Testing.aspx#tab-1

Penuel, W. R., Kell, J., Frost, J., & Khanna, R. (April 1998). *Administrator reasoning about data.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Perie, M., Marion, S., & Gong, B. (2007). *A framework for considering interim assessments*. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from http://www.nciea.org/publications/ConsideringInterimAssess_MAP07.pdf

Perie, M., Park, J., & Klau, K. (2007). *Key elements for educational accountability models*. Washington, DC: Council of Chief State School Officers.

Picus, L. O., Adamson, F., Montague, W., & Owens, M. (2010). *A new conceptual framework for analyzing the costs of performance assessment*. Stanford, CA: Stanford Center for Opportunity Policy in Education, Stanford University. Retrieved from https://scale.stanford.edu/system/files/new-conceptual-framework-analyzing-costs-performance-assessment.pdf

Pizmony-Levy, O., & Green Saraisky, N. (2016). *Who opts out and why? Results from a national survey on opting out of standardized tests*. New York, NY: Teachers College, Columbia University. Retrieved from http://www.tc.columbia.edu/articles/2016/august/results-from-a-national-survey-on-opting-out-of-standardized-tests/

Pogodzinski, J. (2000). *The teacher shortage: Causes and recommendations for change.* San Jose, CA: Faculty Fellows Program, Center for California Studies, California State University.

Polikoff, M. S., McEachin, A. J., Wrabel, S. L., & Duque, M. (2014, January/February). The waive of the future? School accountability in the waiver era. *Educational Researcher, 43*(1), 45–54.

Pollack, J. M., Rock, D. A., & Jenkins, F. (1992, April). *Advantages and disadvantages of constructed-response item formats in large-scale surveys*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Powell, D. L. R. (2015). *Interlude dashboard report: Update on the implementation and impact of Music Makes Us® in Metro Nashville Public Schools.* Nashville, TN: Metro Nashville Public Schools. Retrieved from http://musicmakesus.org/sites/musicmakesus.org/files/mmu_interlude_dashboardreport_1.pdf

Putman, H., Hansen, M., Walsh, K., & Quintero, D. (2016). *High hopes and harsh realities: The real challenges to building a diverse workforce*. Brown Center on Education Policy at Brookings. Retrieved from https://www.brookings.edu/wp-content/uploads/2016/08/browncenter_20160818_teacherdiversityreportpr_hansen.pdf

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioural Statistics, 29*(1), 121–129.

Reback, R., Rockoff, J., & Schwartz, H. L. (2011). *Under press: Job security, resource allocation, and productivity in schools under NCLB.* NBER Working Paper, No. 16745.

Rentner, D. S., Scott, C., Kober, N., Chudowsky, N., Chudowsky, V., Joftus, S., et al. (2006). *From the capital to the classroom: Year 4 of the No Child Left Behind Act*. Washington, DC: Center on Education Policy.

Resnick, L. B. (2006). Making accountability really count. *Educational Measurement: Issues and Practice, 25*(1), 33–37.

Ronfeldt, M. (2015). Field placement schools and instructional effectiveness. *Journal of Teacher Education 66*(4), 304–320.

Ronfeldt, M., & Reininger, M. (2012). More or better student teaching? *Teaching and Teacher Education, 28*(8), 1091–1106.

Ronfeldt, M., Reininger, M., & Kwok, A. (2013). Recruitment of preparation? Investigating the effects of teacher characteristics and student teaching. *Journal of Teacher Education 64*(4), 319–337.

Ronfeldt, M., Schwartz, N., & Jacob, B. (2014). Does preservice preparation matter? Examining old questions in new ways. *Teachers College Record, 116*(10), 1–46.

Sargrad, S., Marchitello, M., & Hanna, R. (2015). *Invisible by design: How Congress risks hiding the performance of disadvantaged students*. Washington, DC: Center for American Progress. Retrieved from https://www.americanprogress.org/issues/education/report/2015/10/29/124421/invisible-by-design/

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation.* Chicago, IL: Consortium on Chicago School Research, University of Chicago Urban Education Institute. Retrieved from http://www.joycefdn.org/assets/1/7/Teacher-Eval-Report-FINAL1.pdf

Schildkamp, K., Lai, M. K., & Earl, L. (2013). *Data-based decision-making in education*. Springer Netherlands.

Schleicher, A. (2015). *Schools for 21st-century learners: Strong leaders, confident teachers, innovative approaches*. Paris, France: OECD. Retrieved from http://istp2015.org/documents/istp2015_oecd-background-report.pdf

Scholes, R. J., & Lain, M. M. (1997, March). *The effect of test preparation on ACT assessment scores*. Paper presented at the annual meeting of the American Education Research Association, Chicago, IL.

Schwartz, H., Hamilton, L. S., Stecher B. M., & Steele, J. L. (2011). *Expanded measures of school performance*. Santa Monica, CA: RAND Corporation.

Shepard, L. A. (2008). Commentary on the National Mathematics Advisory Panel recommendations on assessment. *Educational Researcher, 37*(9), 602–609.

Shepard, L. A. (2010). Next-generation assessments. *Science, 330*(6006), 890.

Southeast Comprehensive Center. (2016). *South Carolina summary report on state report card focus groups*. Cayce, SC: Author.

Stecher, B. M., Epstein, S., Hamilton, L. S., Marsh, J. A., Robyn, A., Sloan McCombs, J., et al. (2008). *Pain and gain: Implementing No Child Left Behind in three states, 2004–2006*. Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/pubs/monographs/MG784.html

Stecher, B. M., & Naftel, S. (2006). *Implementing standards based accountability (ISBA): Introduction to second year findings*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Stecher, B. M., Barron, S., Kaganoff, T., & Goodwin, J. (1998). *The effects of standards-based assessment on classroom practices: Results of the 1996–97 RAND survey of Kentucky teachers of mathematics and writing.* Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing. Retrieved from https://www.cse.ucla.edu/products/reports/TECH482.pdf

Steinberg, M. P., & Garrett, E. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, *38*(2), 293–317.

Steinberg, M. P., & Kraft, M. A. (2016). *The sensitivity of teacher performance ratings to the design of teacher evaluation systems*. Retrieved from http://scholar.harvard.edu/files/mkraft/files/steinberg_kraft_2016_design_of_teacher_eval_systems_wp.pdf

Steinberg, M., & Sartain, L. (Winter 2015). Does better observation make better teachers? *Education Next 5*(1). http://educationnext.org/better-observation-make-better-teachers

Stiggins, R. (2008, September). *Assessment for learning, the achievement gap, and truly effective schools*. Paper presented at the Educational Testing Service and College Board conference, Educational Testing in America: State Assessments, Achievement Gaps, National Policy and Innovations, Washington, DC. Retrieved from https://www.ets.org/Media/Conferences_and_Events/pdf/stiggins.pdf

Strong, M. (2006). *Does new teacher support affect student achievement?* (Research Brief). Santa Cruz, CA: New Teacher Center. Retrieved from https://newteachercenter.org/sites/default/%20files/ntc/main/resources/BRF_DoesNewTeacherSupportAffectStudentAchievement.pdf

Supovitz, J. A., & Klein, V. (2003). *Mapping a course for improved student learning: How innovative schools systematically use student performance data to guide improvement.* Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania Graduate School of Education.

Sutcher, L., Darling-Hammond, L., & Carver-Thomas, D. (2016). *A coming crisis in teaching? Teacher supply, demand, and shortages in the U.S.* Palo Alto, CA: Learning Policy Institute.

Tatsuoka, K. K. (1991). Item construction and psychometric models appropriate for constructed responses. *ETS Research Report Series*, *1991*(2), i-38. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.1991.tb01416.x/abstract

Taylor, E. S., and Tyler, J. H. (2012). *Can teacher evaluation improve teaching?* Education Next. Retrieved from http://educationnext.org/files/ednext_20124_taylortyler.pdf

Taylor, J., Stecher, B., O'Day, J., Naftel, S., & Le Floch, K. C. (2010). *State and local implementation of the No Child Left Behind Act Volume IX—Accountability under NCLB: Final report*. Washington, DC: U.S. Department of Education, Office of Planning Evaluation and Policy Development, Policy and Program Studies Service.

Teoh, M., Coggins, C., Guan, C., & Hiler, T. (2014). *The student and the stopwatch: How much time do American students spend on testing?* Boston, MA: TeachPlus. Retrieved from http://www.teachplus.org/sites/default/files/publication/pdf/the_student_and_the_stopwatch.pdf

Texas Education Agency. (2016). *2016 Accountability Manual.* Retrieved from http://tea.texas.gov/2016accountabilitymanual.aspx

Thapa, A., Cohen, J., Higgins-D'Alessandro, A., & Guffy, S. (2012, August). *School climate research summary* (Issue Brief No. 3). Bronx, NY: National School Climate Center.

Thomas, W. P., & Collier, V. (1997). *School effectiveness for language minority students*. Washington, DC: National Clearinghouse for Bilingual Education

Tooley, M., & Bornfreund, L. (2014). *Skills for success: Supporting and assessment key habits, mindsets and skills in PreK–12.* Washington DC: New America Foundation. Retrieved from http://www.miteacher.org/uploads/1/0/3/4/10347810/11212014_skills_for_success_tooley_bornfreund.pdf

Tyler, U. N., Yzquierdo, Z., Lopez-Reyna, N., & Saunders Flippin, S. (2004). Cultural and linguistic diversity and the special education workforce: A critical overview. *The Journal of Special Education, 38*(1): 22–38.

Urban Teacher Residency United. (2014). *Measuring UTRU Network program impact 2014* [Webpage]. Retrieved from nctresidencies.org/research/measuring-utru-network-program-impact-2014/

Urban Teacher Residency United. (2015). *Clinically oriented teacher preparation*. Chicago, IL: Author. Retrieved from http://nctresidencies.org/wp-content/uploads/2016/01/COTP_Report_Singlepgs_Final.compressed.pdf

U.S. Department of Education. (2012). *Definitions.* Retrieved from http://www.ed.gov/race-top/district-competition/definitions

U.S. Department of Education. (2004). *Innovative pathways to school leadership.* Retrieved from https://www2.ed.gov/admins/recruit/prep/alternative/report.pdf

U.S. Department of Education. (2015). *Fact sheet: Testing action plan*. Retrieved from http://www.ed.gov/news/press-releases/fact-sheet-testing-action-plan.

U.S. Department of Education. (2016a). *Non-regulatory guidance: Using evidence to strengthen education investments*. Washington, DC: Author. Retrieved from http://www2.ed.gov/policy/elsec/leg/essa/guidanceuseseinvestment.pdf

U.S. Department of Education. (2016b). *Non-regulatory guidance for Title II, Part A: Building Systems of Support for Excellent Teaching and Leading.* Washington, DC: Author. Retrieved from http://www2.ed.gov/policy/elsec/leg/essa/essatitleiipartaguidance.pdf

U.S. Department of Education. (2016c). *Public high school graduation rates*. Washington, DC: Author. Retrieved from http://nces.ed.gov/programs/coe/indicator_coi.asp

U.S. Department of Education. (2016d). *The state of racial diversity in the educator workforce.* Retrieved from http://www2.ed.gov/rschstat/eval/highered/racial-diversity/state-racial-diversity-workforce.pdf

U.S. Department of Education, Office of Planning, Evaluation and Policy Development. (2011). *Teachers' ability to use data to inform instruction: Challenges and supports*. Washington, DC: Author. Retrieved from https://www.sri.com/sites/default/files/publications/teachers_ability_to_use_data_to_inform_instruction_challenges_and_supports.pdf

U.S. Department of Health and Human Services. (2010). *Association between school-based physical activity, including physical education and academic performance*. Retrieved from http://www.cdc.gov/healthyschools/health_and_academics/pdf/pa-pe_paper.pdf

Ushomirsky, N., Hall, D., & Haycock, K. (2011). *Getting it right: Crafting federal accountability for higher student performance and a stronger America*. Washington, DC: The Education Trust. Retrieved from http://edtrust.org/wp-content/uploads/2013/10/Getting_It_Right.pdf

Ushomirsky, N., Williams, D., & Hall, D. (2014). *Making sure all children matter: Getting school accountability signals right.* Washington, DC: The Education Trust.

Villar, A., & Strong, M. (2007). Is mentoring worth the money? A benefit-cost analysis and five-year rate of return of a comprehensive mentoring program for beginning teachers. *Educational Research Service, (25)*3, 1–17.

Virginia Department of Education. (2015). *Statistics & reports: Accreditation & federal reports*. Richmond, VA: Author. Retrieved from http://www.doe.virginia.gov/statistics_reports/accreditation_federal_reports/

Voight, A., Austin, G., & Hanson, T. (2013). *A climate for academic success: How school climate distinguishes schools that are beating the achievement odds* (Report Summary). San Francisco, CA: WestEd. Retrieved from http://files.eric.ed.gov/fulltext/ED559741.pdf

Wagner, L. (February 4, 2015). *Do A–F school grades measure progress or punish the poor?* NC Policy Watch. Retrieved from http://www.ncpolicywatch.com/2015/02/04/do-a-f-school-grades-measure-progress-or-punish-the-poor/

Wagner, L. (2016). *Do A–F grades measure progress or punish the poor?* Raleigh, NC: NC Policy Watch. Retrieved from http://www.ncpolicywatch.com/2015/02/04/do-a-f-school-grades-measure-progress-or-punish-the-poor/

Wallace Foundation. (2016). *Improving university principal preparation programs.* Retrieved from http://www.wallacefoundation.org/knowledge-center/Documents/Improving-University-Principal-Preparation-Programs.pdf

West, M. R. (2016). *Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California's CORE districts.* Washington DC: Brookings Institute. Retrieved from https://www.brookings.edu/research/should-non-cognitive-skills-be-included-in-school-accountability-systems-preliminary-evidence-from-californias-core-districts/

West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F., et al. (2015). Promise and paradox measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis*, 0162373715597298.

West Virginia Department of Education. (2016). *West Virginia's A–F school grading system FAQs*. Retrieved from https://static.k12.wv.us/a-f/a-f_faqs_flyer.pdf

Whitaker, S. D. (2000). Mentoring beginning special education teachers and the relationship to attrition. *Exceptional Children, 66*(4), 546–566.

Whitaker, I. (2016). *Fruitful recruiting: CCSD makes strides in addressing teacher shortage.* Las Vegas Sun. Retrieved from http://lasvegassun.com/news/2016/jul/13/fruitful-recruiting-ccsd-makes-strides-in-addressi/

Whitehurst, G. J. (Russ), Chingos, M. M., & Lindquist, K. M. (2014, May). *Evaluating teachers with classroom observations: Lessons learned in four districts*. Washington, DC: Brown Center on Education Policy, Brookings Institution. Retrieved from http://www.brookings.edu/~/media/research/files/reports/2014/05/13-teacher-evaluation/evaluating-teachers-with-classroom-observations.pdf

Wilgoren, J. (June 9, 2001). *Possible cheating scandal is investigated in Michigan*. New York Times. Retrieved from http://www.nytimes.com/2001/06/09/us/possible-cheating-scandal-is-investigated-in-michigan.html

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181–208.

Wise, L. L. (2011, March). *Picking up the pieces: Aggregating results from through-course assessments*. Paper presented at the invitational research symposium on Through-Course Summative Assessments, Atlanta, GA. Retrieved from http://www.ets.org/Media/Research/pdf/TCSA_Symposium_Final_Paper_Wise.pdf

Wong, M., Cook, T. D., & Steiner, P. M. (2009). *No child left behind: An interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series.* Evanston, IL: Institute for Policy Research.

Wynn, S. R., Carboni, L. W., & Patall, E. A. (2007). Beginning teachers' perceptions of mentoring, climate, and leadership: Promoting retention through a learning communities perspective. *Leadership and Policy in Schools*, *6*(3), 209–229.

Appendix A: List of Stakeholders

Assessment

Pedro Anes
President
Inpact Learning, LLC

Carol Barone-Martin
Executive Director, Early Childhood Programs
Pittsburgh Public Schools

Jill Bartoli
Associate Professor of Education
Elizabethtown College

David E. Baugh
Superintendent of Schools
Centennial School District

Dennis Baughman
Principal
York Academy Regional Charter School

Joan Benso
President and CEO
Pennsylvania Partnerships for Children

Elizabeth Bolden
President and CEO
Pennsylvania Commission for Community Colleges

Victoria Rice Campbell
Transition Consultant, Special Education Teacher
Allegheny Intermediate Unit

Joseph Cannella
Supervisor of Data and Achievement
Radnor Township School District

Jackie Cullen
Executive Director
PA Association of Career and Technical Administrators

Mark DiRocco
Superintendent
Lewisburg Area School District

Ted Grice
Math and Science Teacher and Math Department Chair
Belle Vernon Area School District

Otis Hackney, III
Chief Education Officer
City of Philadelphia

Jane Hershberger
Supervisor of Multilingual and Multicultural Supports
Chester County Intermediate Unit

Rachel Holler
Director of Programs and Services
Bucks County Intermediate Unit

Julie Huff
Assistant Superintendent for Academics
Mechanicsburg Area School District

Jackie Mills
Math Teacher
Philipsburg-Osceola Area School District

Christine Oldham
Superintendent
Ligonier Valley School District

Christopher Shaffer
Deputy Chief, Curriculum, Instruction, and Assessment
School District of Philadelphia

Misty Slavic
Director, Curriculum and Instruction
Freedom Area School District

Diane Wilkin
Visual Arts Educator, Bristol Township School District;
President, Pennsylvania Art Education Association

Accountability

Andrew Coonradt
Instructional Initiatives Coordinator
Delaware County Intermediate Unit

Donna Cooper
Executive Director
Public Citizens for Children and Youth

Sharif El-Mekki
Principal
Mastery Charter School–Shoemaker Campus

Michael Faccinetto
2016 Vice President,
PA School Boards Association;
President, Bethlehem Area School Board

Brad Ferko
Superintendent
Sharpsville Area School District

Daniel Fogarty
Director of Workforce Development
Berks County

Richard Fry
Superintendent
Big Spring School District

Susan Gobreski
Director for Community Schools
City of Philadelphia

Carey Harris
Executive Director
A+ Schools

Lisa Harris
2nd Grade Teacher
Woodland Hills School District

Angela King
Counselor
Chester County Technical College High School

Cheryl Kleiman
Staff Attorney
Education Law Center

Mark Korcinsky
Principal
Seneca Valley School District

David Lapp
Director of Policy Research
Research for Action

Stacey Marten
Math Teacher
Hempfield School District

Amy Morton
Chief Academic Officer
Central Susquehanna Intermediate Unit

Lensi Nikolov
Director of Instructional Planning and Monitoring
Allentown School District

Bryan O'Black
Assistant Superintendent
Shaler Area School District

Matthew Patterson
Director of Elementary Education
Corry Area School District

Stephen Shaud
Director
Elwyn SEEDS

Stinson Stroup
Manager, Education Services
PA State Education Association

Educator Preparation

Jodi Askins
Executive Director
PennAEYC
Rhonda Brunner
Assistant Executive Director
Capital Area Intermediate Unit

Allison Burrell
District Librarian and Media Specialist
Southern Columbia School District

Joan Duvall-Flynn
President
NAACP Pennsylvania State Conference

Julie Fabie
Director of Early Childhood Education
York City School District

John Friend
Superintendent
Carlisle Area School District

Joel Geary
Certification Officer
Penn State Harrisburg

Tracy Hack
PECT/Praxis Coordinator
Butler County Community College

Mark Holman
Director of Human Resources
School District of Lancaster

Amy Lightner
Data and Instruction Specialist
Central Dauphin School District

Catherine Lobaugh
Assistant Executive Director for Early Childhood, Family, and Community Services
Allegheny Intermediate Unit

Meredith Mehra
Director of Professional Development
School District of Philadelphia
Judy Morgitan
School Nurse
Perkiomen Valley School District

Gwyneth Price
Department of Middle and Secondary Education and Education Leadership, School of Education
Edinboro University

Katherine Rutledge Speech and Language Pathologist
Bucks County Intermediate Unit

Abby Smith
Director of Education
Team PA Foundation

Linda Torres
UniServ Representative
PA State Education Association

Sarah Ulrich
Program Director, Teacher Education
Drexel University

Sally Winterton
President
PA Association of Colleges and Teacher Educators

Educator Evaluation

Marnie Aylesworth
Early Childhood Director
The Pennsylvania Key

Shandia Booker
Counselor
Pittsburgh Public Schools

Korri Brown
Special Education Teacher
Kennett Consolidated School District

David Christopher
Superintendent
Juniata Valley School District

Cathleen J. Cubelic
Director, Curriculum, Instruction, and Assessment
Midwestern Intermediate Unit

Carolyn Dumaresq
Retired Educator
Former Secretary of Education

Sam Franklin
Professor
Carnegie Mellon University

Michelle Harris
Professional Learning Specialist
School District of Philadelphia

Paul Healey
Executive Director
PA Association of Elementary and Secondary School Principals

Tara Huber
English Teacher
Neshaminy School District

Allison Mackley
Teacher-Librarian
Derry Township School District

Cindy Minnich
English Teacher
Upper Dauphin Area School District

Jerry Oleksiak
President
PA State Education Association

Larry Redding
Superintendent
Gettysburg Area School District

Denise Rogers
Middle Representative
Philadelphia Federation of Teachers

Sherri Smith
Superintendent
Lower Dauphin School District

Kathy Swope
2016 President, PA School Boards Association; President, Lewisburg Area School Board

Teresa Szumigala
Principal
Erie School District

Louise Tharp
Certified School Nurse
Warren County School District

Tim Wagner
Associate Principal
Upper St. Clair School District

Lynette Waller
Director of Elementary and Secondary Education Wyomissing Area School District

Appendix B: Links to Additional Resources

For summary notes, background research, and other materials associated with the April 28, June 14, and August 30 meetings of the ESSA work groups, visit the Pennsylvania Department of Education's ESSA website at http://www.education.pa.gov/Pages/Every-Student-Succeeds-Act.aspx.

ABOUT AMERICAN INSTITUTES FOR RESEARCH

Established in 1946, with headquarters in Washington, D.C., American Institutes for Research (AIR) is an independent, nonpartisan, not-for-profit organization that conducts behavioral and social science research and delivers technical assistance both domestically and internationally. As one of the largest behavioral and social science research organizations in the world, AIR is committed to empowering communities and institutions with innovative solutions to the most critical challenges in education, health, workforce, and international development.

AIR®

AMERICAN INSTITUTES FOR RESEARCH®

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000

www.air.org

Making Research Relevant

LOCATIONS

Domestic

Washington, D.C.
Atlanta, GA
Austin, TX
Baltimore, MD
Cayce, SC
Chapel Hill, NC
Chicago, IL
Columbus, OH
Frederick, MD
Honolulu, HI
Indianapolis, IN
Metairie, LA
Naperville, IL
New York, NY
Rockville, MD
Sacramento, CA
San Mateo, CA
Waltham, MA

International

Egypt
Honduras
Ivory Coast
Kyrgyzstan
Liberia
Tajikistan
Zambia

7333_10/16