Teacher Professional Evaluation and Professional Growth Systems in Maine: 2015 Report

Prepared by:

Craig A. Mason, Ph.D.

Shihfen Tu, Ph.D.

June 2015

Maine Education Policy Research Institute
College of Education and Human Development
University of Maine
Orono, Maine

Teacher Professional Evaluation and Professional Growth Systems in Maine: 2015 Report

June 2015

Craig A. Mason, Ph.D.
Professor of Education and Applied Quantitative Methods

Shihfen Tu, Ph.D.
Associate Professor of Education and Applied Quantitative Methods

Maine Education Policy Research Institute
A nonpartisan research institute funded by the Maine State Legislature, the University of Maine, and the University of Southern Maine.

Center for Research and Evaluation
College of Education and Human Development
University of Maine, 5766 Shibles Hall, Orono, Maine 04469-5766
(207) 581-2493

A Member of the University of Maine System

AUTHOR'S BIOGRAPHICAL INFORMATION

Craig A. Mason, Ph.D., is a Professor of Education and Applied Quantitative Methods at the University of Maine, where he also serves as the Director of the Center for Research and Evaluation and Co-Director of the Maine Education Policy Research Institute. For the past decade, Dr. Mason has also served as a methodological consultant to the U.S. Centers for Disease Control and Prevention. Dr. Mason received his PhD in Clinical Child Psychology from the University of Washington, and his research interests are in developmental growth models, parent-child relationships, informatics, and research methods. He has over 80 publications, and has been principal investigator or co-principal investigator on over $10 million in grants.

Shihfen Tu, Ph.D., is an Associate Professor of Education and Applied Quantitative Methods in the Department of Exercise Science and STEM Education at the University of Maine, where she is also affiliated with the Center for Research and Evaluation (CRE) and MEPRI. Dr. Tu has extensive experience in overseeing statewide projects in child health and development. She is the PI of a project that developed and currently maintains a statewide database system, *ChildLINK*, which manages early childhood screening data from programs in the Maine Center for Disease Control and Prevention (Maine CDC), Children with Special Health Needs (CSHN) Program. She is also involved in a longitudinal data analysis project with the Maine Department of Education State Longitudinal Data System.

EXECUTIVE SUMMARY

At the request of the Maine State Legislature, the *Maine Educational Policy Research Institute* (MEPRI) has monitored the progress and challenges school districts have faced in designing and implementing teacher performance evaluation/professional growth (PE/PG) systems. During the last year, particular attention has focused on federal requirements that PE/PG systems use ***statewide*** standardized assessment data for measuring student growth over time. Therefore, MEPRI conducted a series of case studies involving seven school districts across the state in order to assess issues involving the incorporation of student growth data in their PE/PG system. The work for this project was conducted in Spring of 2015, with the goal of addressing three general sets of questions:

(1) What instruments are these districts currently using to assess student growth across the curriculum? What features do superintendents and teachers seek in student assessment measures? To what degree are districts using the MEA/Smarter Balanced for assessing student growth, and what concerns do they have that may be limiting its use in PE/PG systems?

(2) How are these districts using student data to define and measure growth? How is growth weighted and incorporated into their PE/PG system?

(3) What classroom observation tools are these districts using? What challenges and solutions have they found? How do student growth and classroom observation data compare? How are they balanced and reconciled in the PE/PG system?

This project involved interviews with superintendents and/or their designees for seven school districts in Maine. Districts were selected based on information provided in previous MEPRI surveys in order to identify those that (1) were using standardized student assessment data to measure student growth, (2) were at a relatively more advanced stage of PE/PG system design and implementation, and (3) reflected a range of student, community, and geographic variation across the state of Maine. Case studies used a semi-structured interview conducted in-person or via a conference call. Interviews were recorded with the permission of all participants, and subsequently transcribed.

Student Assessment Instruments

Not surprisingly, one of the more significant challenges districts have faced is identifying reliable, valid, standardized instruments for assessing student growth across the broad array of academic content areas covered in K-12 education. Researching and selecting specific instruments often involved a complex and time consuming process conducted by teams of faculty and administrators.

Beyond validity, reliability, and alignment with the curriculum, superintendents noted several additional key features that were considered when selecting measures of student academic growth. First, the time and scheduling of the assessment were key considerations: an assessment should be efficient, meaning it provides a maximum amount of useable information, but requires a minimal disruption to the normal classroom schedule and practice. Second, superintendents felt assessments should include tools that allow teachers to identify individual student strengths, weaknesses, and learning-gaps in as much depth as possible. This information could then be used by teachers to develop a more student-centered, individualized curriculum. Third, several superintendents were specifically interested in measures that allowed multiple assessments each year. This would provide a more complex and sophisticated view of student growth, but would also potentially allow them to be used as a formative assessment tool. Similarly, superintendents were interested in measures that would provide results back to administrators as a way of informing district policy and actions. Finally, several superintendents specifically noted a desire to avoid instruments that created an environment where there were opportunities for manipulation – or even for the potential appearance of manipulation – by educators. This was specifically in response to recent action in Atlanta where several educators were sentenced to jail for illegally manipulating student assessment data.

Beyond mathematics and reading/writing, superintendents indicated that their districts continued to seek standardized measures in other content areas. For example, some districts are using *Fit Stats* – a physical performance assessment already used by many schools in Maine—as a measure of growth in physical education, while others are addressing growth in the performance arts through change in portfolios or common performances over time. To identify measures, some districts have drawn upon instruments developed and used by multiple different sources and state PE/PG systems. Superintendents are also leveraging other existing or upcoming

assessment efforts, such as RTI and proficiency-based education, as a way to address PE/PG assessment needs. Partnerships with other districts, as well as guidance from the Department of Education, can thus be particularly beneficial in identifying such solutions. This can help districts avoid "recreating the wheel", as well as potentially help offset the cost of researching and implementing solutions.

Superintendents expressed a number of concerns with incorporating the MEA / Smarter Balanced assessment into their PE/PG systems. Foremost was the fundamental question of whether the Smarter Balanced assessment was going to be used beyond its first year. Districts were uncomfortable shaping the student-growth portion of the PE/PG system around a tool that may only be in place once. The degree to which superintendents felt that key information regarding Smarter Balanced results continued to be unclear was an additional major concern. These included alignment with curriculum and future policy, such as proficiency-based education, as well as the type, extent, and format in which the results would be provided to educators. Also, districts that had institutionalized a formal vetting process for reviewing and selecting instruments were hesitant to adopt a new tool without applying the same standards and review process to a new MEA measure. While Maine has since decided to discontinue using Smarter Balanced, these concerns may prove valuable when selecting a new measure.

Superintendents observed that the alignment of assessments with coursework will require particular attention over the coming years. The transition to standards-based education may lead to significant changes in the content and timing of some material. Districts will need to monitor their curriculum and assessment instruments in order to verify that what is being measured actually aligns with what is intended to be covered in the classroom.

Incorporating Student Assessment Data into PE/PG Systems

With no officially defined formula or approach for translating student assessment data into measures of student growth, districts have adopted a range of strategies. Depending on the instrument used and the number of assessments, growth was measured within a single academic year (fall 2014 to spring 2015), across one calendar year (spring 2014 to spring 2015), or across several years (spring 2011-spring 2015). Some districts also used different time scales based on the specific course, with growth over a single year applied to courses that are covered on a

regular, steady basis (e.g., mathematics), and growth over multiple years used for courses that have more limited instructional time (e.g., performance arts). Districts also varied in how they addressed summer learning loss. Spring-to-spring assessments include any loss that occurs over the summer: The greater the summer learning loss, the greater the improvement needed to "break even" with the previous end-of-year spring assessment. However, as noted by one superintendent, ignoring summer learning loss may overstate how much ***true*** growth is occurring over multiple years and unintentionally lead to schools not exploring solutions for this issue.

Districts included in these case studies generally weigh student growth as 20% or more of a teacher's PE/PG score, or are building to 20% over the next few years. Individual student growth was aggregated at different levels in different districts. For example, the student growth score for a sixth grade Spanish teacher may reflect the sum of (1) growth observed for students in her class, and/or (2) the overall growth for ***all sixth grade Spanish students*** in the school, and/or (3) the overall growth for ***all sixth grade students*** in the school, and/or (4) the overall growth for ***all students*** in the school. These districts may have the growth component for a teacher's PE/PG score based on all four of these different levels of aggregation, or for educators who do not work with a specific class, their PE/PG score may be weighted differently to focus on student growth in the program area (e.g., Spanish) or grade. Superintendents reported that with good communication obtaining support for a final formula for student growth was generally achievable with few difficulties.

One of the challenges with incorporating student performance data into a PE/PG system is identifying an official teacher of record. Superintendents reported that it was important to include a degree of flexibility in assigning teacher of record in order to address unique situations that may arise, while recognizing the need to carefully evaluate and monitor such exclusions in order maintain the validity of the entire system and avoid "cherry-picking" student scores. A deeper concern was that too much attention on the teacher of record may lead educators to focus solely on children who they perceive as "their" students, at the expense of providing support and assistance to other students around them. This was particularly true in regards to students in special education, where several districts relied heavily on co-teaching.

Observation and Other Student PE/PG Data

Not unexpectedly, observations of teachers "in action" in their classroom were uniformly seen as vital components to assessing the quality teaching. Districts use a variety of classroom observational tools based on different standards or models of teaching, with superintendents uniformly satisfied with whatever specific system their district was using. Observational systems were computer/web-based, with a range of tools to help observers make reliable, accurate assessments. Systems also included reporting tools to assist teachers and supervisors interpret the results and identify skill-areas in which a teacher may benefit from further attention and training. Some districts also include peer observations conducted by other teachers. These may be formal or informal, and depending upon the district they may be strictly for a teacher's own edification and not made available to administrators as part of PE/PG evaluations. In particular, peer observations were seen as a tool for encouraging discussion, collaboration, and idea-sharing among teachers.

Unfortunately, the costs for implementing an observational system and training observers to reliability can be significant. This can be offset in part through partnerships with other districts, although superintendents uniformly reported that support from the state for observations would be valuable. In particular, lack of state funding was seen as potentially placing smaller districts at a relative disadvantage. Other possible state-level support, such as state-sponsored professional development or regional training, or state assistance coordinating larger collaboratives would also be appreciated by the districts.

Some districts also incorporate student surveys into their PE/PG system. Depending upon the district, student surveys may be widely implemented or used on a limited scale in response to specific concerns, such as contradictory PE/PG data. Superintendents noted that while teachers may initially be uncomfortable with the idea of being "evaluated" by their own students, the information was valuable in providing insight into the student perception of the classroom experience.

Ultimately, while some teachers were uncomfortable with PE/PG systems identifying teachers as performing at different levels of effectiveness, superintendents also reported that other teachers were positive about different levels of performance being recognized. Particularly hard-working

and high-performing teachers may be frustrated by a system that simply places all teachers into the same category. Furthermore, it is difficult to target and address the need for additional training and support if the evaluation system fails to flag those teachers in need of such support.

Making it Work: Superintendent Suggestions

Finally, superintendents noted several common strategies they felt were valuable when developing and implementing their systems. These include:

- Start early and meet target dates. However, superintendents also expressed frustration that in doing so they had to make repeated changes as state or federal guidelines changed.

- Draw from multiple sources of information. This leads to more reliable and valid summaries of teacher performance and effectiveness, and serves to address teacher concerns regarding potential problems or biases with any single source.

- Throughout the design process, meet regularly in order to maintain momentum. Working with other districts is an effective way to share ideas and leverage resources.

- An open, inclusive membership in the design process is valuable.

- Clear, regular communication with teachers and administrators is important in order to ensure transparency as well as to identify and correct any misconceptions that may arise.

- Everyone must see PE/PG as a continually ongoing process, not just a "hoop to jump through" every few years. All educators should be engaged in some type of PE/PG activity each year. As goals are met, new ones should be established.

- The PE/PG process cannot simply be seen as a punitive tool used to discipline teachers. The goal is to help improve teaching for ***all*** educators, and thus improve student learning. A perception that the system is designed to target poor teachers interferes with it being used to help promote better teaching.

TABLE OF CONTENTS

INTRODUCTION

In April 2012, LD1858 was signed into law, setting Maine on a path to develop comprehensive teacher performance evaluation / professional growth (PE/PG) systems, with the goal of enhancing educator effectiveness and student learning and achievement in Maine. At the request of the Maine State Legislature, the *Maine Educational Policy Research Institute* (MEPRI) has monitored this process through annual surveys of Maine superintendents. These surveys have shown how the PE/PG process for various districts has progressed at different rates based on a number of factors, including access to resources (e.g., a *Teacher Incentive Fund* or TIF grant), existing local assessment practices, staffing changes, and local motivations and concerns.

Furthermore, as the development, piloting, and implementation process has unfolded, state and federal rules and legislative action have resulted in changes to the timeline and requirements for PE/PG systems. During the last year, particular attention has focused on federal requirements that PE/PG systems incorporate not just standardized assessments of student growth, but specifically ***statewide*** standardized assessment data. For Maine, this is the *Maine Education Assessment* or MEA. In Maine, this change was further complicated by the state's transition from the *New England Common Assessment Program* (NECAP) to the *Smarter Balanced* assessment as the MEA tool effective Spring 2015.

Therefore, MEPRI conducted a series of case studies involving seven school districts across the state in order to assess issues districts have faced and strategies they have developed for including student growth data in their PE/PG system. For context and comparison, these case studies also examined the use of classroom observational data in connection to how these two very different approaches are balanced and potentially reconciled in the PE/PG process. Districts were identified based on their previous annual PE/PG survey reports and selected so that the case studies focused on those that were at a relatively more advanced stage of the design and implementation process, particularly in regards to the use of student assessment data.

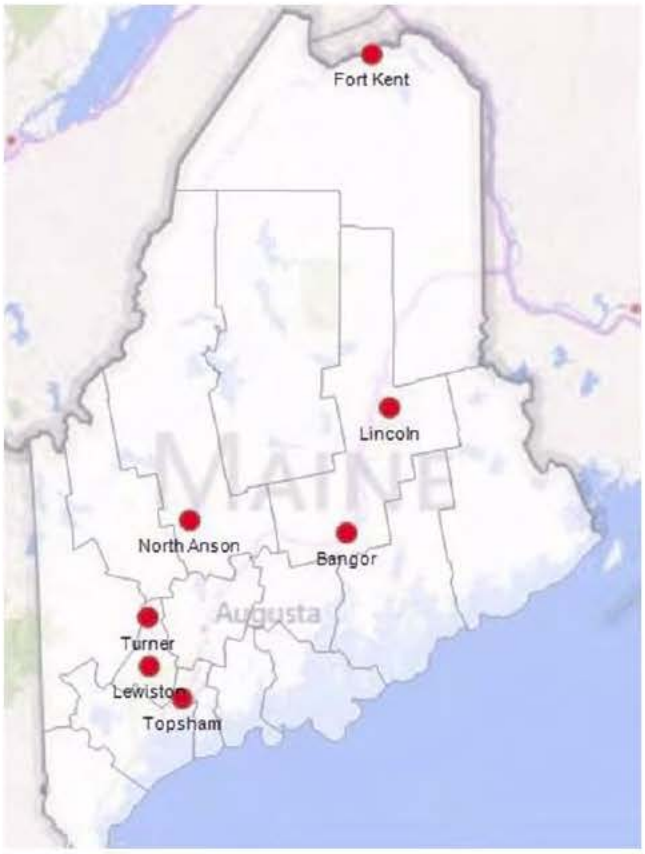
The work for this project was conducted in the Spring of 2015, with the goal of addressing three general sets of questions:

(4) What instruments are these districts currently using to assess student growth across the curriculum? What features do superintendents and teachers seek in student assessment measures? To what degree are districts using the MEA/Smarter Balanced for assessing student growth, and what concerns do they have that may be limiting its use in PE/PG systems?

(5) How are these districts using student data to define and measure growth? How is growth weighted and incorporated into their PE/PG system?

(6) What classroom observation tools are these districts using? What challenges and solutions have they found? How do student growth and classroom observation data compare? How are they balanced and reconciled in the PE/PG system?

Figure 1. Participating Districts



This report integrates the findings from these seven case studies. It begins by discussing the selection of student assessment instruments, including the MEA, and how districts are addressing assessment across the curriculum. It then describes how student assessment data is being used to define and measure growth over time, and how growth is used in the PE/PG system. The report then reviews classroom observation data and observational systems being used by these districts, including the use of student surveys of the classroom environment. The report then concludes with superintendent observations and suggestions for designing and implementing a PE/PG system.

METHODS

This project involved interviews with superintendents and/or their designees for seven school districts in Maine. Districts were selected based on information provided in previous MEPRI surveys in order to identify those that (1) were using standardized student assessment data to measure student growth over time, (2) were at a relatively more advanced stage of PE/PG system design and implementation, and (3) reflected a range of student, community, and geographic variability across the state of Maine.

Specific districts included in these seven case studies were:

- Bangor
- Lewiston
- MSAD 27 (Fort Kent)
- RSU 52 (Turner)
- RSU 67 (Lincoln)
- RSU 74 (North Anson)
- RSU 75 (Topsham)

Case studies were conducted during Spring 2015 using a semi-structured interview of the district superintendent and/or their designee. On average, interviews lasted approximately 1 hour and occurred in-person or via a conference call. Interviews were recorded with the permission of all participants, and subsequently transcribed. Direct quotes from the interviews are used throughout the report. Verbatim quotes are used whenever practical. Square brackets, or [ ], are used when the author has paraphrased spoken material and braces, or{ }, are used to reflect other information that may be useful when interpreting the speaker's voice, such as {laughter}.

All participants gave permission to be identified and for quotes to be used in this report; however, speakers are nevertheless de-identified (i.e., Superintendent A through G, based upon a random ordering). In some cases the source for a quote may simply be labeled "a superintendent" if the information is likely to be identifiable and thus also reveal other quotes he or she has made. Nevertheless, given the characteristics of these districts and Maine, it is likely

that the source for many specific quotes can be determined if a reader so desires; however, as noted previously, participants gave permission to be quoted.

When reading these quotes, it is important to remember that these are verbatim statements of unprepared spoken material, largely left unedited in order to capture the speakers “voice”. The flow and grammar thus reflect a more casual and colloquial style than the speakers would use in prepared text or speech.

Superintendent B: It's important for people to understand what a standardized assessment is, and that does not necessarily mean an Iowa or a MEA or an NWEA. And that non-standardized assessments, by their very nature, are extremely problematic in terms of utilization for equitable and comparable assessment of teacher effectiveness. I would love someone to explain to me how [non-standardized assessments] could be utilized in an equitable and comparable system that provided equitable comparability of performance assessment across multiple teachers. I'd love to have that explained to me if someone believes it can be done….

Not all standardized assessments are created equal, and not all come from commercial entities. [There] seems to be a misconception that a standardized assessment is a commercial assessment from some company far away. I hope that [districts and the state] will be able to look at the assessment itself on its merits, rather than on who happened to be the origin of it.

STUDENT ASSESSMENT INSTRUMENTS

The section summarizes various measures and strategies that participating districts are using to assess student performance across the curriculum. It includes areas such as mathematics and reading that have a more established history of standardized assessment, as well as historically less-addressed topics such as performance art and physical education. The MEA/Smarter Balanced assessment is specifically addressed, including concerns districts reported with adding Smarter Balanced (or any new MEA assessment) into the PE/PG process. The section ends with a discussion of features and characteristics districts seek when choosing assessment measures.

SELECTION OF ASSESSMENT INSTRUMENTS

Not surprisingly, one of the more significant challenges districts have faced is identifying reliable, valid, standardized instruments for assessing student growth across the broad array of academic content areas covered in K-12 education. Researching and selecting specific instruments often involved a thorough and complex process of

(1) Reviewing measures

(2) Comparing and contrasting administration features (e.g., time demands, formatting),

(3) Evaluating reporting options (e.g., breadth/depth of information provided to teachers)

(4) Considering the timeliness with which schools would obtain the results,

(5) Determining the alignment to state and local curriculum guidelines.

For districts in these case studies, this work was conducted by teams of faculty and administrators committing considerable time and energy to the process.

> *Superintendent B: We use a wide range of student assessment data, based on the developmental level and the instructional content of the teacher in question. The first issue that we encountered was making sure that…the assessment actually was measuring learning outcomes which were previously [taught] during the instructional cycle. So the selection of assessment tools was the first and most critical choice that had to be made in that process. To do that, we gathered a team of faculty and administrators to review assessments that were currently in place, assessments that were readily available, and assessments that could be created to address that need to measure student learning outcomes in the learning standards that were within the instructional cycle of the teacher... The team went through those broad categories of potential assessments and selected assessments that were considered appropriate by the team, within the criteria outlined in the statute and rule.*
>
> *Superintendent D: We have a grid of different ELA, math, science, and social studies assessments that are common to all in a grade level. Those are the ones that we're using. If you started in pre-K it would be "Land of the Letter People." It would be the early math diagnostic assessment, the early literacy assessment. If you go up to middle school, we have our document-based questions as local assessments in social studies. [We also] incorporate national assessment, state assessment, and local assessment. We had been using the NECAPs and the TerraNovas. However, we're kind of treading water right now because with Smarter Balanced we left the TerraNova.*
>
> *Superintendent B: [For] mathematics and English language arts [this includes] Northwest Evaluation Associates, Measures of Academic Progress. Also included, in the area of science, are discipline-specific NWEA assessments... In addition, we utilize for primary students, the DIBBLES assessments….*

As this suggests, for some districts, the result was a rather extensive list of instruments that were incorporated into a comprehensive assessment system, with different tools used based on grade-level and academic area.

Drawing on a large pool of instruments, one strategy used by some districts allows teachers or schools to select specific instruments from a portfolio of possible measures. Teachers are then able to measure growth for different students using different instruments in order to address the unique needs or specific concerns of a child.

> *Superintendent A: We call them our options. It's sort of like a drop-down menu that could be NWEAs,… classroom scores, we use the STAR math in some grade levels, STAR reading, even Smarter Balanced once that it's taken could be something that would be in the dropdown menu for teachers [to use when assessing growth].*

NON-MATHEMATICS/READING CONTENT AREAS

Nevertheless, superintendents indicated that their districts continued to seek instruments – particularly in areas outside of math, reading, and writing. Widely adapted, standardized measures are less common for some content areas, and so superintendents reported that partnerships with other districts, as well as guidance from the Department of Education, can be particularly beneficial in identifying such tools. By sharing information regarding options and experiences, districts can avoid "recreating the wheel" or worse, repeating mistakes. Furthermore, partnerships can help to offset the cost of researching and implementing possible solutions.

> *Superintendent C: We haven't [addressed many areas outside of math and reading/writing] yet. We are part of the collaborative, the Maine Cohort for Customized Learning, so we're still hammering out getting all the pieces [of] the Maine Learning Results into "Empower". We just had this long discussion this morning and we're getting our arms around getting all that data. MCCL re-did the literacy pieces, so they've modified all the stuff that goes in "Empower." That's going to help, but what do we do about the PE teachers and the health teachers and how's that data going to fit in there? I don't know yet.*

> *Superintendent A: We haven't invested in the science or social studies yet because they're a little afraid of the amount of time the testing would take away from the direct instruction. There's only so much testing we can put kids through.*

One area where there exists regular opportunity to measure student growth, but fewer well-known standardized assessment tools is Physical Education. Some districts have addressed this using *Fit Stats* – a physical performance assessment already used by many schools in Maine.

> *Superintendent B: In the area of physical education, we feel we're doing pretty well. We're using performance -- student outcomes in terms of fitness levels – [through] a commercial product called Fit Stats, which provides… the standardized assessment… with developmental norms across different age levels, and allows us to look for growth in the levels of fitness of our students across time, using a common measure. That's been very helpful in that area.*

Similarly, performance arts and related disciplines are other areas in which some districts find it difficult to incorporate measures of student growth. However, they continue to identify options, even if some of these are based on more limited standardization and historical validity.

> *Superintendent A: If I walk into art class and perform a self-portrait of myself on the first week of school, and I do it again on the last week of the class, have I gained? Have I gotten better? Is my writing improving? [Maybe we need] examples in a portfolio that shows that it has over time. I'm not sure if the sit-down "test kids to death" [approach] is the best way. These are all things that need to be tracked in some type of student portfolio.*

> *Superintendent B: Additional assessments that have been selected are [instruments] that come from the assessment bank in Washington state, including an assessment called "Get the Part", which is used as a pre- and post-assessment for performing arts. Also from that assessment bank, [we use] music content area assessments and visual art assessments. For visual arts, we are utilizing a rubric-based assessment of a performance task. The performance task is the creation of a visual art product with the target the… creation of a still life piece by the students….In foreign language, we're utilizing an oral performance assessment, developed by the foreign language association of Maine… Each of the student performances is double scored by [outside] individuals who have been trained in the rubric, and the mean of the double scores is used as the final result, unless there's significant discrepancy in the scores, indicating a lack of reliability. [In that situation] we have a third piece that comes in as a juried process to establish more accurately what the score is. We use that same process in the performing arts, the visual arts, and music – where the scorers do not know the identity of the individuals whose work is being scored. Nor do they have a direct connection to the instructor whose students are being assessed.*

This latter district has drawn upon instruments developed and used by multiple different sources in order to maximize the number of instructional areas addressed by their student assessments. For example, their measures for secondary-level social studies / history were taken from work in New York.

> *Superintendent B: …We also have adopted the New York Regents Exams, to be used both as pre-assessments and post-assessments for selected coursework in social studies and humanities at the secondary level -- specifically, US History, World History.*

Perhaps the most challenging area for assessment in connection with PE/PG systems involves growth among students receiving special education services. Depending upon the unique status of a child, the range of possible skill areas and developmental levels that may be targeted by educators at any grade level can be significant. The process if further complicated by other state initiatives, such as proficiency-based education, that have their own impact on assessment. The result is that many districts continue to explore strategies for assessing students in special education, while holding true to the expectation of continual improvement over time.

> *Superintendent E: We're still grappling with life skills---what would be the best assessment to use with some of those students that maybe are, for example, non-verbal?*

> *Superintendent E: [When measuring growth] students with disabilities currently have the same expectations… What I keep saying is we don't want them to be losing more than a year's growth. So if you go into it with the mindset of "oh they shouldn't be growing as much as another student", then they're just going to keep getting further and further behind. And this is where we merge the proficiency-based diploma, and talk about if we go into it with that mindset, those children will never be able to graduate with a proficiency-based diploma. So even though they may have a disability, we can make modifications, we can make accommodations, and we have a place where they can note that.*

Understandably, assessment-focused monitoring and reporting on student growth may be particularly new and stressful for many educators in certain areas. However, with sufficient support, information, and training, superintendents reported most teachers found the result positive for themselves and their students.

> *Superintendent G: And while there was initially angst among our unified arts teachers, as we've gotten more and more into it, there are things that we can and should be measuring. I think it's created some great conversation and great movement in terms of what are our expectations for these non-tested subject areas.*

ADDITIONAL POSSIBLE STRATEGIES FOR COLLECTING STUDENT DATA

In some cases, districts are simultaneously in the process of adapting their curriculum in ways that may facilitate assessing student growth in courses outside of mathematics/reading/writing. This is not a case of "teaching to the test", but rather incorporating relevant, interdisciplinary skill-building based on reading/writing and mathematics into other content areas. For example, one superintendent described how in order to improve student writing skills, their district previously adopted a policy that involves mandatory writing across all disciplines. This led to a "Reading, Writing, Running" course, in which students wrote about their physical activity. The school similarly developed a writing-cooking class where students wrote about their experience preparing, serving, and sharing meals. As this superintendent noted:

> *Superintendent A: We went to mandatory writing across all the disciplines. You had to do writing in physical education, you had to be writing in English, you had to be writing in social studies, and science. It wasn't just left up to one area. Let's say that you're social studies. We don't have a formal assessment like we do in reading and in math, but as a social studies teacher, you have a choice of taking the NWEA reading scores, and doing something with that… [But, how do you] use the reading piece of the NWEA, even though you're a social studies teacher or science teacher? That's something that probably most [teachers] haven't looked at or seen before…. So it's going to be an adjustment to say,… "this is something that we feel all kids [need to do] -- you need to be a good reader to be a good social studies student." And how much of your class involves reading? I think a lot of them do. It might be that I need to get better at understanding reading comprehension and how to teach comprehension in my social studies class. [It will involve] more than just assigning "section two of chapter four" for tomorrow's homework. That's not going to cut it. They're going to have to become a teacher of reading to [some] degree. If you're going to give a reading assignment what should it look like, what should it entail, how rich is it? Is it just something students have to do? I don't want to pick on social studies but if you remember the old social studies textbooks, they weren't the best piece of reading. There are primary documents out there that can be used and that are a lot more entertaining and engaging than just a history textbook.*

As described by this superintendent, doing so may require significant work on the part of teachers, but result in a more positive and engaging educational experience for the students.

Superintendents are also leveraging other assessment efforts that either exist in their district or are on the horizon, such as response-to-intervention (RTI) or proficiency-based education. One

noted how their district incorporated their existing RTI screening tool (Aimsweb) for literacy and math into their local assessment system.

> *Superintendent F: This is where two of our large initiatives overlap, in that we're also working on our transition to proficiency-based [education]. There's the recognition that we're going to need to have assessments to measure students' proficiency anyway, so if we don't already have something, [a new measure] could really be used for two different things.*
>
> *Superintendent F: There are some assessments that are still being utilized as an outcome of the local assessment system years back. That's primarily at the high school level, where those may be resurrected. Or if they're still being used they would be used for this purpose now, where they may not have been before.*

Leveraging existing assessments may be an efficient way to collect PE/PG data without placing additional testing demands on students and teachers. Nevertheless, instruments should not be adopted simply as a matter of convenience. When conducting these reviews, superintendents reported that in many cases it became apparent that an existing measure may work well for assessing student learning, but not actually address the need to assess the impact of ***teaching*** on student learning.

> *Superintendent B: We [selected instruments] through a collaborative process, first looking at what assessment tools we are currently [using] to measure student progress and for determining instructional placement… The general theory was that if we're measuring the learning standards with this tool for the student, logically that would be the measure of whether the teacher was effective in implementing instruction that supported the student in meeting those learning standards. So our first step was to examine the assessments that we utilized for student assessment purposes and see if they would fit with the criteria of the rules and the law and be valid or reasonable in measuring teacher effectiveness. We found that some tools that we were using seemed to be, and some tools were more problematic -- even though they served a legitimate purpose in the student assessment process, particularly in a formative fashion, they were less helpful being used for measuring the impact of teacher instruction towards the attainment of the learning standards. One assessment that we had been trying to use, but found to be too problematic to continue was the diagnostic reading assessment.*

ASSESSING AREAS BEYOND ACADEMICS

With the focus on student academic growth, several superintendents expressed concern that assessments may be overlooking growth in valuable non-academic skill areas. "21st Century Skills" such as perseverance, creativity, and problem solving, are increasingly recognized as important skills not generally addressed in traditional standardized assessments. Moreover, a system that is seen as focusing on a few high stakes test results may discourage students from exploring new areas or from taking healthy academic risks which require them to push their limits.

> *Superintendent A: I think there was a need for some testing, just to keep everybody honest. But… as we move into the realm of content is important but it's not the only thing anymore. [Being a] lifelong learner, the problem-solver, the communicator… It's getting kids to understand that perseverance, grit, is very important in the in larger world. But if you fail something, it doesn't mean you quit. It means you take it over and get better at it. And I don't find that [as much] in today's teenagers. I find a lot of them will get their first F at [UMaine] in accounting and change majors. You know? And no, it's okay to get an F in accounting -- it's a tough subject, and you need to take it over. And when you take it over you get an A and you're going to be okay.*

> *Superintendent: If you want to come live and work here [outside of the larger cities in Maine], you might need skills that aren't directly correlated to a four year degree. You may not need a four year degree. But you may need some college, some two-year degree, a lot of work experience -- someone who knows the importance of coming to a company, working, and staying.*

THE MEA / SMARTER BALANCED

Not surprisingly, several of the participating superintendents did report using the MEA for student growth measurements.

> *Superintendent E: I feel very comfortable using either one of those [the MEA or NWEA]. In some places we are using both of them, especially in the math area.*

> *Superintendent F: [Our choice to use the MEA] was recognition of the fact that there were assessments already in place for measuring both literacy and math, and the ability to have some reliable data around those two.*

Nevertheless, even with the easy access to the MEA as a standardized instrument, these superintendents expressed significant concerns with incorporating it into their PE/PG systems.

Foremost was the fundamental question of whether the Smarter Balanced assessment was going to be used beyond its first year. Districts were uncomfortable shaping the student-growth portion of the PE/PG system around a tool that may only be in place once.

> *Superintendent D: However, we're kind of treading water right now because with Smarter Balanced we left the TerraNova. We thought of going to NWEA, but then the state was trying to convince us that we didn't need NWEA because Smarter Balanced would have interim assessments and pre-assessments. But [from what I am hearing] I believe you're going to see that they'll leave Smarter Balanced, and I think we could be headed towards [some other instruments]. We would love that to get settled because this is very unlike [our district] to be treading water. We don't like that feeling because that's not who we are. So we need the state to make a decision and then we'll determine how to best round out our assessment data.*
>
> *Superintendent G: And now we have the MEA Smarter Balanced, and you know, we'll see where that ends up.*
>
> *Superintendent B: One of the questions previous to now was whether [the Smarter Balanced Spring 2015 assessment] would actually happen. As late as January, there were still questions in our mind whether the technology would actually function in a way that would allow the assessment to be administered. Now, that's proved to be not as bad as it could have been. But we're still not sure exactly what will happen in terms of processing the results and reporting. We have to look at all that to make a decision based on how well it will meet the purpose, and how well it can serve -- legitimately and accurately -- the purpose of assessing educator effectiveness in impacting student attainment of the learning standards.*

Beyond the fundamental question of whether Maine would continue to use Smarter Balanced, districts were reluctant to transition their PE/PG systems to rely on an instrument for which they felt there were significant unknowns. One district that ***was*** planning to use Smarter Balanced as a measure of student growth was potentially interested in using fall interim assessments as a pre-test measure. However, at the time of the interview, this decision was on hold pending further information regarding the nature and scaling of the interim assessments.

> *Superintendent F: We know that [the pre-assessment] has to be before any kind of impact that the teacher has had. When it comes to MEA it could be that we use the spring to spring -- you know, the end of 3rd grade for instance to the end of 4th grade -- to be that measure. But we've also learned…that some of the interim assessments are also aligned on the same scale. So we have to learn*

> *more…. We have to wait to hear more from the state about the interim assessments. Could there be an interim assessment that's given in the fall, before the teacher has the influence and then use the summative at the end of the year [as] the post? It's kind of like we have some missing information that we need from the state to know more about the interims before we can really land on whether it's going to be within the same school year, or if we're comparing the end of 3rd grade to the end of 4th grade, for instance.*

On a deeper level, questions regarding Smarter Balanced also touched on the basic psychometric qualities and alignment with curriculum and policy, such as proficiency-based education. Superintendents also reported that the type, extent, and format in which the results would be provided to educators remained unclear. Districts that had institutionalized a formal vetting process for reviewing and selecting instruments were thus hesitant to adopt a new tool without applying the same standards and review process.

> *Superintendent G: Well I think that to some extent it may be unfair to compare a NWEA assessment that's been refined over a number of years with the first administration of the new MEA assessment. You know, we know that the MEA is, while adaptive, not to the same extent the NWEA was. The MEA as an interim tool is also in its infancy. Then you've got the whole Common Core movement around it, which is very politicized right now, and whether there's going to be some something definitive out of Washington remains to be seen, but probably unlikely….So just a lot unknowns there.*

> *Superintendent C: We haven't seen the Smarter Balanced outcomes yet. So, we don't even really know where to go with it until we see what they're going to look like.*

> *Superintendent B: We take a kind of a pragmatic approach to defiance {laughter}. We did not exclude the MEA for categorical reasons, but rather because we didn't have enough knowledge of it previously to make a determination. So one of the things we'll be doing this summer and next year is to look at how that rule can best be responded to. One of the things that I understand is [that] the state MEA for certain grade levels must be incorporated in the performance assessment of teachers who have those students… in the content areas being assessed. But that will be a decision that the whole group will look at and make based on new knowledge about the assessment, now that we've had it administered at least once. And there'll be more data and information that we can analyze, both in terms of its structure and its reporting forms.*

This touches on a more fundamental issue in which districts that began the process early and/or worked aggressively to meet state deadlines have systems that are relatively well-developed. Many of these districts "did the right thing" and implemented successful policies, procedures, and practice, and are now understandably cautious about potentially upsetting what they have built – particularly given concerns (at the time) regarding the future of Smarter Balanced and the impact other state initiatives.

> *Superintendent G: But we have our student learning objectives (SLOs), and we've made great progress with that. Our steering committee has yet to approve anything, and we might well end up with two SLOs or an SLO and the MEA. One of the challenges in framing the MEA, however, is we still don't know how the results are going to look. By that I mean the format of the results – the reporting – and just how you would frame an SLO using that. That should be clearer over the next few weeks, but remains one of our challenges.*

> *Superintendent E: With NWEA you get the results when you leave the screen, and then you can dig right down into it. It's a quick kind of dipstick, where we can use it midyear for kids who are at risk. Right now it's the only thing that we have that's been continuous. When they keep changing the test, it makes it tricky. So if they'd quit changing the test, I guess we may [use the MEA and drop NWEA] if it really becomes more timely, which I don't know whether the MEA will be. Right now we give the NWEA at the beginning of the year. For the kids who are at risk, many of them take it in the middle of the year, and then we give it again in May. And you're not going to be able to give the MEA like that. So right now, I don't see us getting rid of that NWEA in the next couple of years.*

Districts reported varying levels of support for incorporating a new MEA measure, in large part depending upon how it compares to their existing instruments and procedures.

> *Superintendent B: If the collaborative group determines that in a specific case the MEA assessment is superior to the assessment that we have available in its accuracy and appropriateness in measuring teacher impact on student growth and student attainment of learning standards, then I'm confident they will propose the replacement of that prior assessment with this as the new assessment. However, if their determination is that it's not superior – that the existing assessment is superior in achieving that, then my belief is they will minimize the use of it. If it's comparable, then my guess is that they will choose to use it as an additional measure in a blended approach. So it really depends not on a pre-determination or a pre-judgment of the appropriateness of the assessment, but putting [the MEA] through the same kind of review and critique*

that we put all the other assessments through in terms of… measuring student learning and growth.

An option noted by one superintendent would be to include the MEA if required, but if it does not satisfy their district needs or criteria, to continue with their existing measures and simply apply a minimal weight to the MEA contribution in the final score.

Superintendent B: That being said, it's helpful that at this time there has not been any specific weighting attached to the inclusion of [the MEA]. So theoretically, as we read it right now, at the extremes one could say that we have three measures of student learning and growth, and one of them is one of the content areas in the MEA. [One non-MEA] measure will constitute 49.75 % of the contribution to the teacher's student learning and growth measure. The [other non-MEA measure] will be 49.75% and the MEA will be 0.5 % of the contribution.

Finally, it is worth noting that while superintendents some felt that recent concerns regarding the amount of testing students experience were at times overstated, others reported that they felt the time required for the Smarter Balanced assessment was problematic. Beyond the time students spent on their own assessment, these superintendents reported that the scheduling and coordination of assessments had a larger negative impact on the school as a whole. For example, any high school course that covered several grade levels was potentially impacted by the 11th grade assessment.

Superintendent E: I think the whole thing about over-testing our kids is exaggerated. The NWEA basically takes an hour for the math, an hour for the English. And the MEA definitely takes a little bit longer, but when you talk to teachers you'll hear them say things like "oh we've been testing for a month." I think that's slightly exaggerated {laughter}. The test window might be a month. We're not testing any one student for a month.

Superintendent A: How much testing do we really want? If you look at our Junior year, it is amazing how much testing they have to do. We actually showed that to the school board the other night and we broke out all the grade levels and the testing that came up. We found out our 11th graders go through quite a bit of testing, and then our seniors go through none. It's just interesting.

Superintendent D: Personally, I'm not opposed to the questions in Smarter Balanced, but the testing window and the implementation process is a nightmare. For that reason, I would be more in favor of going to NWEA. We're going to lose children, in my opinion, because you cannot test from the beginning of March to the end of May in this methodology. They're exhausted. It is

impacting their school schedule. No one likes it. It's just too long. It's overload. And I can't even imagine doing interim assessments. It is much better to have a quick hit, if you will {laugh}. Come in, two weeks, out. NWEA I think can do that, but Smarter Balanced can't.

While many of these points are now moot given the state decision to leave Smarter Balanced, the implications that such change and action have on district perspectives and potential future behavior is worth considering as Maine goes on to select a replacement.

DESIRED PROPERTIES OF STUDENT ASSESSMENT INSTRUMENTS

Beyond validity, reliability, and alignment with the curriculum, superintendents noted several additional features that were considered when selecting measures of student academic growth. While a decision regarding the future use of the Smarter Balanced assessment had not been made at the time interviews were conducted, the perspectives offered by these superintendents may prove valuable in identify future state-wide assessment tools.

The time and scheduling of the assessment were key considerations. Superintendents reported weighing the amount and usefulness of the information obtained from different instruments against the time required to administer them. In essence, an assessment should be efficient, meaning it provides a maximum amount of useable information, but requires a minimal disruption to the normal classroom schedule and practice.

Superintendent C: We are using STAR. STAR is a ____ product. I saw it in action [at another district] and it's a pretty powerful tool if you use it to leverage raising the bar for your kids, about getting an idea of where they are as proficient or non-proficient.

Superintendent G: You know, we had the NECAP. Of course that was given at the wrong time of year.

Superintendent D: We would prefer spring assessment rather than October {note: this was in reference to the NECAP}. I think that the October window disrupts the start of a healthy school year. We made it work, but it was so much calmer this year, and probably more for the adults than the children. When adults feel like they've covered their routines at the beginning of the year and that students are adhering to those expectations, it's just better for everybody.

Superintendent C: I know there's a lot of data that you can get from [Smarter Balanced], but I'm seeing that STAR gives me a heck of a lot of data too on a 15,

> *20 minute assessment. [Plus] I can give it as many times as I want, and it really zeroes in an RTI for the kids. So that's where we are right now. Smarter Balanced is what, a one-shot deal, six hours? I haven't seen any data. I haven't seen any outcomes from it. So that's why I thought we've got to do something.*

Continuing with this line of reasoning, superintendents also felt assessments should include tools that allow teachers to identify individual student strengths, weaknesses, and learning-gaps in as much depth as possible. This likely involves user-friendly online tools or detailed student-level teacher reports. It was felt that this information should be "vertically stratified" so that beyond simply identifying that a student is performing above or below expected levels for a given grade, interested teachers can easily determine the grade-level at which a student is performing and the specific content knowledge he or she has or has not yet mastered.

> *Superintendent E: One of the things that I love about the NWEA is that it's helping to reinforce [the goal] that all of our students should be making growth. And it's not just about teaching to the Common Core standard, because of the way the results are on the NWEA you can really drill down and figure out what they need to be taught. I think that's kind of a paradigm shift for some of our teachers.*

To fully-leverage student assessment data, this information would also involve teachers expanding their pedagogical skills and curriculum plans in order to potentially incorporate material outside of what is typically expected for a given grade-level. This may require significant changes to the classroom, but would reflect a more comprehensive transition to a student-centered, individualized curriculum.

> *Superintendent E: [Historically, teachers would often] feel like if I'm teaching 4th grade I need to know the 4th grade curriculum. They haven't thought deeply enough [that] even though you are teaching 4th grade, you may have kids at the 2nd grade level, or you may have kids at the 6th grade level. And you need to figure out what is the next step for them to continue to move forward. So that's one of the things I really like about the NWEA.*

> *Superintendent A: I think we need to use the results to better plan our instructional approaches in our classrooms and what curriculum is needed…I think we do need periodic testing, but… we also need to see what that data [is] telling us about a school. If we're 38% meeting or exceeding in reading and writing, then maybe the structure of the school day next year looks a little different. We add more of that for kids. [We don't want to] just think what we're doing is working and it's just a bad score…*

> *Superintendent A: [When it comes to individual students who are struggling]-- he's not just going to get better. She's not [just] going to get better. Over time I think to jump up a grade level is [maybe] a half hour extra a day in either math or English [or whatever topic] you're doing poorly in.*

As this suggests, for many teachers the use of assessment data to directly inform practice may require additional training and professional development. In some districts, it may also require providing teachers with additional student-level information in user-friendly formats. But the end result is a more data-informed educational practice that leverages student assessment data in order to individualize teaching and learning.

> *Superintendent E: We have some training scheduled for this summer about how to dig down into the NWEA data to help inform your instruction. Because that's one of the things we've said from the beginning. I said we should not be creating any assessment for the purposes of teacher evaluation that doesn't help inform our instruction. If we're doing that, it's not a good assessment… Data's wonderful, but it doesn't mean anything unless you know what to do with it. And I think that's the piece that's often left out when we give assessments, any kind of assessment. You generate this data and you can say we got this number, but so what? What do you do next?*

> *Superintendent A: We want [teachers] to look at previous student data. So if I'm a 9th grade English teacher, I need to look at the 8th grade student data, for each kid, and how they did in 8th grade. I need to set my lesson plans up, and I need to steer my work around that incoming student data. [I] no longer write this generic lesson plan – I need to write a lesson plan that is going to meet their needs, because of what the data say.*

Related to this, several superintendents were specifically interested in measures that allowed multiple assessments each year. This would provide a more complex and sophisticated view of student growth, but would also potentially allow them to be used as a formative assessment tool. One superintendent noted a particular strength of their current student-learning measure was the ability for teachers to use it for ongoing formative assessments.

> *Superintendent C: If I listen to the folks that are saying the Smarter Balanced assessments are going to do all these neat and fabulous things, then maybe STAR goes away. But I hope not because it's a 15, 20 minute assessment that I can do formatively throughout the year. I don't know what we're going to see with Smarter Balanced. I can't give six and a half hour exams once a month {laugh}. But that's the value, that's how I'm selling the school board and the community on the investment in technology. Because if we have these things in the kids'*

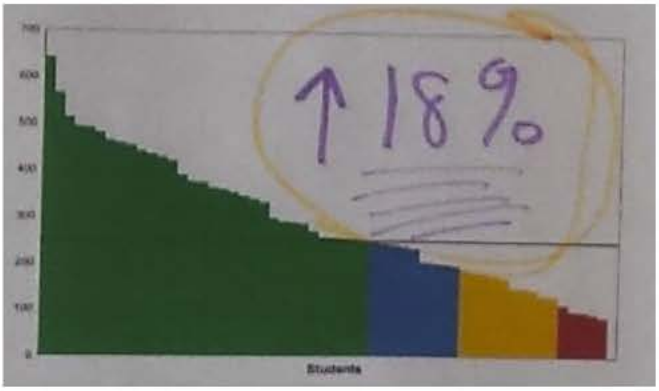
hands, we can really leverage those things to do these kind of assessments… [and research shows] you get the biggest bang for the buck providing formative evaluation to the teachers.

When done right, formative assessment is not seen as an extraneous "check" on students and teachers, but rather it becomes a seamless valuable core part of the iterative teaching process.

Beyond teacher's using student data to inform their own practice in the classroom, superintendents were also interested in using data to help inform district policy and actions. For example, one superintendent described sharing aggregated grade-level data generated by their system with teachers, school board members, and other key constituents as a way of informing them of areas of strength and concern when shaping local educational policy and practice.

Superintendent C: And I actually have [STAR results] posted in the conference room, and anybody that comes in, [such as] school board members, I take them in there and I show them. I say, "Look, here's where we are. This is the gap. This is why I have this item in the budget." So I'm using the STAR data for that right now. But the teachers also know it's on the wall down in the superintendent's office. [They know] he's looking at it… {Note: See Figure 2 for an example of the visual display illustrating increased proficiency rates at one grade level}

Figure 2. Illustration of Student Proficiency Rates



Regardless of their final selection of instruments, several superintendents specifically pointed to recent events in Atlanta, where educators were sentenced to jail for racketeering in connection with the manipulation of student testing data. This concern is reflected in conscious decisions to avoid options that create an environment where there were opportunities for manipulation – or even for the potential appearance of manipulation.

Superintendent D: [We] never, never [want to put] the teacher at risk of the potential of what looks like cheating. I've told them we want a system that protects you and protects the system. So there is a way that we are looking at it together and it's not one teacher on their own saying "See, I have notable achievement gains" and yet they're the only ones who scored or looked at the data. Because that's really dangerous.

Superintendent B: One of the issues that we had, in terms of our overall process, was to avoid Atlanta [laughter]. So one of the things that I believe is that those kinds of things happen when the people designing the system don't take into account the way human beings actually behave, but rather, [think in terms of] the way they would like them to behave. One of the precepts that we tried to do is to avoid the opportunity for [manipulating] outcomes, or, there's an old Arabic saying: "The man who leaves his horse untied in the street is as guilty as the thief." So we don't leave the horse in the street. One of the problems with the DRA is the nature of it as a formative assessment is that the teacher in question is the person administering and scoring and evaluating the assessment. We found that incorporating that into the process would place the teacher, at best, in a position of temptation and potential suspicion. At worst, in a position of opportunity to manipulate the results. We didn't want to put our teachers in that circumstance.

INCORPORATING STUDENT ASSESSMENT DATA INTO PE/PG SYSTEMS

The prior section focused on measures of student performance across the curriculum and strategies that districts were employing in order to address historically less-assessed content areas. This section reviews how participating districts are incorporating these measures into their PE/PG system. It includes a summary of how districts conceptualize and operationalize "growth" using multiple assessments over time, as well as the relative weight that different districts apply to the student growth component of their systems. The section concludes with a discussion of how factors outside of the classroom can negatively impact student growth and teacher PE/PG scores.

HOW IS GROWTH CONCEPTUALIZED AND CALCULATED

With no officially defined formula or approach for translating student assessment data into measures of student growth, districts have adopted a range of strategies, from strictly quantifiable, to quantifiable with a large visual/qualitative component to help interpret the results, to a more holistic focus.

> *Superintendent C: The whole district is [assessed using STAR] in October, then again when we came back from the break, in January, and again in June. But [teachers] can give it at any time. Some of the teachers are really pushing the envelope, and they're the ones that are seeing the most growth. They're really looking at the data. They're giving them monthly, almost as a formative assessment to figure out where the weak points are….In fact, in a discussion we had this morning, the special education [teachers] at the high school really want to start digging into it so they can get their arms around where their kids are.*

> *Superintendent E: We're going with half the gap. At first [the committee] wanted to go with* **any** *growth, and then we had a lot of discussion about if the kids only grow one point, they're going to keep getting further and further behind. So the committee [is] on board with upping the standard for all of our kids. [Related to this], we have had discussions about where are our kids [are in terms of] graduating with proficiency. So we keep trying to link those two. They are separate, but in many ways they're integrated. I've stressed that if we're not having adequate growth, then there is no way some of these kids will be able to graduate with a proficiency-based diploma. So we merge the two when we can.*

Superintendent E: [This is then applied to] the same model that the state is using for the report card grade. So if you have a student who scored in the bottom half of 1, then the next year, for that to be adequate growth, we would want them to come up to at least the middle half, or if they're a 4 that they're maintaining a 4. Or if they're a 3, they're either maintaining or going up to the 4.

Superintendent A: We have our math and reading boards and we code kids in colors. Red meaning that you're two or three years below [expected level], yellow meaning that you're either one year behind or on target, and then green means that you're either on target or two or three years above…. So we code each [student] by color, with their student id number, not their name. We start the year out with the most current data that we have. From there we meet about every month and after every formal assessment that we do, like the NWEA or Smarter Balanced, and we re-code them on the board. Each grade level teacher [then] goes in and meets for an hour and a half to talk about what they'll do for each kid, or the class in general…. Some kids have started out in the red and now are up in the green levels, and we ask ourselves "why the great gains?" And [when a student declines] we also ask ourselves why did that kid drop two tiers this year? What happened? Was it an outside factor? Was it my instruction? Why, what happened with this kid? Was there something that happened in their life? At their home? Or here at school?

Beyond how growth is measured, the timing of the assessments fundamentally shapes how growth is conceptualized and what extraneous factors may impact the results. For example, annual spring-to-spring assessments inherently incorporate some degree of "summer learning loss" into student's growth and PE/PG results: The greater the summer learning loss, the more disadvantaged the teacher will be in the following year (i.e., the greater the improvement needed just to "break even" with the previous end-of-year spring assessment). However, as noted by one superintendent, ignoring summer learning loss may overstate how much ***true*** growth is occurring over multiple years. Rather than recognizing and addressing an issue impacting student learning, it essentially removes a school's responsibility for finding innovative solutions for summer learning loss.

Superintendent G: The NWEA we potentially administer three times a year. This year we made the January one optional at the individual teacher's decision. In terms of the assessment data, and this I think is an important point, for this NWEA we did a spring to spring measurement. Summer learning loss is particularly large in a community with a population [such as ours]. So if we only measure fall to spring, we, as a school district are not taking any responsibility for involvement in what happens during the summer. And I maintain that a day

> *in July is as important as a day in January, from the student perspective. So that's why we did spring to spring. Because otherwise you'd have teachers showing success but not seeing it in the students. Our typical student loses the equivalent of two months of instruction if they don't go to summer school. If they go to summer school, generally speaking, they're able to maintain. That's just some of our own internal research, but that's also in what we've read nationally.*

As noted previously assessing student performance in certain content areas is challenging and requires more innovative solutions. The same is true for using the results of those assessments to measure growth over multiple years/assessments. In performance arts, one superintendent noted that their district evaluated growth across a three-year time period in order to account for more limited instructional time.

> *Superintendent B: Students complete these still life works annually, and the prior year still life is utilized as the pre-assessment. The end-of-year still life is used as the post-assessment. So, the post-assessment for one year becomes the pre-assessment for the following year. Because of the limited instructional time in the visual arts area, we measure growth over a three year time span, with a staggered process. So first grade students' product in first grade becomes the pre-assessment for the fourth grade post-assessment. With the second, third, and fourth years being the instructional cycle—the three year period. The end of second grade work becomes the pre-assessment for the fifth grade post-assessment, et cetera, et cetera. That allows us to provide what the team felt was equalized opportunity for the teacher to impact the student development in the learning targets, despite the diminished annual instructional time.*

This same superintendent also noted challenges addressing industrial arts, with current plans focusing on assessing growth in industry certification-related skills.

> *Superintendent B: We haven't been able to find similar tools to measure growth in learning standards in the area of industrial arts. That's been probably one of the most difficult areas to find any assessments that fit. At the present time, we're working on developing a methodology that relies on assessing students meeting external industry certification standards. Sort of a movement from students having not met the standards necessary for certification in a specific trade area, to having met some or all of the certification requirements in a specific trade area, such as masonry or carpentry. [We would then look] for growth in the number, percentage, or level of those industry certification standards that students have met. That's the approach we think is going to be workable and appropriate for the content area.*

As noted previously in connection with student assessment data, ideally evaluating student growth will occur across multiple independent assessments and instruments, rather than a single high stakes test.

> *Superintendent D: Let's say one of our reading goals is to focus on informational texts. We would want notable gains as demonstrated through the assessments [using] informational texts. We would pull the informational text assessments and analyze that data [from various assessment instruments, such as] the TerraNova…Fountas and Pinnell… or the SRI. And our district writing assessment would [also] be on informational texts. So we would look at our improvement through a number of assessments rather than one high stakes assessment.*

The challenge becomes collecting these multiple measures in a way that is as seamless and unobtrusive as possible for the regular classroom teaching environment.

WEIGHT APPLIED TO STUDENT GROWTH

As state policy has evolved, there has been considerable debate regarding how much weight student growth data should carry with a district's PE/PG system. Districts included in these case studies have adopted various models, generally weighing student growth as 20% or more of a teacher's PE/PG score, or building to 20% over the next few years. Some districts don't assign a specific percentage to various components of the PE/PG system, but include student growth as a core part of a larger matrix.

Districts also differ in how they aggregate student growth data. For example, the student growth score for a sixth grade Spanish teacher may reflect the sum of (1) growth observed for students in her class, and/or (2) the overall growth for ***all sixth grade Spanish students*** in the school, and/or (3) the overall growth for ***all sixth grade students*** in the school, and/or (4) the overall growth for ***all students*** in the school. In some districts, the growth component for a teacher's PE/PG score may be based only on the classroom-level aggregation, in others it may be based on all four aggregation levels. Alternatively, if she does not work with a specific class, it may be weighted to focus on student growth in her program area or grade.

> *Superintendent A: {Currently], we are going to start with 10% and work our way up over the next two or three years at a greater margin. So maybe the next year it would be 15% and then the following year it would be the 20% that the state is*

look for. That is a really thin piece of ice for teachers. They're really scared of that.

Superintendent G: We have a hundred point scale. We have 78 points on fifteen professional standards, of which the majority are observable in the classroom, and the remaining ones are where [a teacher could provide] additional evidence or [based on] observations outside the classroom. [These include] how a teacher relates to and collaborates with their peers….Then we've got 5 points on the whole school measures, which at this point has been the NWEA, 15 points on student learning objectives and 2 points on peer observation and collaboration, for a total of a hundred. [The student learning objective results] are based upon the percentage of students who met the growth goals set for them at the beginning of the year. And it's a direct, zero to 15 points, spread out from 20% or less meeting that goal to 96% to 100% meeting that goal. There may be some changes in that going forward, but at this point in time that's what we've done.

Superintendent F: Because [our system is] aligned to this matrix idea, we haven't said 80% for this and 20% for that. It's just that these components will match up onto different axes on a matrix. Obviously the student growth is going to have a bit more of an impact because it's only being matched up on a matrix to the combined results, from the [other areas]. So there is more of an impact from that than maybe the [other components] would have. But we haven't necessarily articulated specific weights yet to each of those components.

Superintendent E: We went with 20%. [Teachers outside of math/reading] will have teacher-created assessments around their subject areas. For teachers that don't have any student data, we changed the proportions of the rubrics for determining teacher effectiveness. So a regular teacher that would have student data, their professional instructional skills are 60%, their goal-setting is 20%, and student data is 20%. But for teachers that do not have that student data piece, their professional practice is 70% and their goals is 30%. Initially [percentage for student growth] was 20%, and then there was some leeway given.

Ultimately, participating superintendents reported that with good communication and understanding by all parties, obtaining support for a final formula for student growth was generally achievable with few difficulties. This was particularly true for some districts that had been engaged in the process longer, in some cases pre-dating official state efforts.

Superintendent E: I think teachers looked at [using 20% for growth] and felt that it wasn't something that was unreasonable because we kept reminding them, "look you've got 60% in your professional practice, which you have control over, and another 20% in your goals." So if you've got 80% you're going to be an effective person, even without student data. So I think that was a piece.

Superintendent B: We had the advantage that the team that was working on this had been working on this since before the educator effectiveness law was actually passed. Because this is actually our 5th year of being involved in the process, they had a different view of that issue than has been prominently raised across the state. And their decision was a blend that was 50/30/20, with 50% of the weight being on the professional practice standards, including the combination of supervisor observation and portfolio, 30% based on student learning and growth measures, and 20% based on student perception surveys … And they were most concerned about the accuracy and reliability of the supervisor observation rating – more so than the accuracy and reliability of the student learning and growth measures or the student perception surveys. So they balanced the traditional approach against their concerns, by saying that the supervisor portion should not be more than 50% of the final summative rating.

Superintendent B: [In a referendum on the process and initial plan] out of 60 teachers, 54 cast ballots, and of those 52 cast ballots in support of the collaborative team continuing with the program they had and the process that had been developed, and continuing to refine and move forward from that point, including the 50-30-20 plan. Now, I don't take that as meaning that every teacher who voted in favor of the team was in favor of every element within the plan. But it provides an overall vote of support and confidence for the general direction, and particularly for the blend, because that was such a highlighted and prominent piece of the system that was put forward to them.

TEACHER OF RECORD

One of the challenges with incorporating student performance data into a PE/PG system is identifying an official teacher of record for students. Several superintendents described various criteria their district considered for determining the teacher of record. Inevitably, they identified scenarios where certain students would "slip through the cracks", creating a situation where no one was responsible for some youth. For example, students who receive special education, ELA, and/or Title I services in schools with team teaching may fall into a category where no single educator satisfies some predefined threshold of contact required of the teacher of record.

Superintendent D: You don't get to say these are my children and those are your children. They're all our children.

Superintendents reported that finding solutions for this problem led to valuable conversations among faculty and administrators in many of these districts. These discussions, combined with state changes designed to help resolve such issues, resulted in changes in some districts that reflect the broader goal of helping children regardless of traditional structures and barriers.

Superintendent F: As we are making changes to how we approach supporting students in a proficiency-based system, [we see that there may be] multiple teachers that are teaching [a given] fourth grader math, based on how we're being flexible in addressing the needs of that student. Some of the language was changed in the version of the rule that we have now, that I think would support it more than the way it was written before. Again, it allowed it before, but I think it acknowledges it a little bit better now.

Superintendent G: I think it's resulted in some good conversation about who ultimately needs to take responsibility for that child. And it should be the classroom teacher. But if the special education teacher or the ed tech is doing less than the classroom teacher feels they should be, should the classroom teacher [have] the authority to say "I can do better with that student"? I mean I think a case could be made that this could contribute to less team teaching or sharing of students, just because it's so difficult to break out those measurements.

Even with this reflection and review, superintendents reported that deriving a single definition or algorithm that would identify the appropriate teacher of record in all cases may not be possible without the flexibility to address unique situations that may arise. Consequently, districts have incorporated varying degrees of flexibility in how the teacher of record is identified. For example, one solution provides teachers the option – with principal approval – of having specific student data removed from their PE/PG results based on factors such as student mobility. However, superintendents also recognized the need to carefully evaluate and monitor such exclusions in order maintain the validity of the entire system and avoid "cherry-picking" student scores.

Superintendent G: We give the teacher the benefit of the doubt. Obviously a student that's there for the pre-assessment and for the post-assessment is part of that cohort that the teacher's going to be judged upon. But in terms of those students that come in late, or leave early, or are pulled out, teachers can propose -- and with the principal's authority -- students can be pulled from their score. [It requires] flexibility.

Superintendent F: We were careful to make sure that we weren't putting in lots of exclusions in how a teacher defined who their cohort was going to be. In some ways at the high school and middle school-level for instance, whoever's on your roster is essentially who your cohort is. So you couldn't cherry-pick, so to speak, who's going to be in the cohort, because obviously that's going to have an impact on their growth.

> *Superintendent E: At the high school, if I'm teaching a course…those kids that I have [are the ones for which] I'm the teacher of record. If I'm teaching history I'm the teacher of record for those kids for history…. We're doing their primary courses, not their electives, and at the high school level it's primarily in terms of the student data, [which are] teacher-developed instruments at this point. So whatever the topic is -- math, science, English, PE -- it really matches the class the student is in….*

Superintendents also expressed concern that an unintended consequence may be less collaborative partnership among educators in the teaching of individual children. The fear was that it may lead educators to focus solely on children who they perceive as "their" students, at the expense of providing support and assistance to other students around them.

> *Superintendent D: I really think that whoever you're sharing this with needs to hear this: We have developed a culture where they're all our children. You don't get to say these are my children and those are your children. They're all our children. And by having a team approach, it is very common to go to a school where teachers are working together to figure out who can do more reading with a student at need, because we do data walls, and we analyze the data through a visual representation of students, by the whole school, of who's moving and who's not. And it's very common to hear the PE teacher say "Listen, I've got 20 minutes right here. I'll sit and read with a child." And if you don't have goals and measurement that honors all of them together, I think you get more of what you see across the nation, where teachers are complaining about which teachers get which kids. And I want to do whatever I can to avoid that. It's not about pitting adults against adults. They're all our kids.*

SPECIAL EDUCATION

Issues of student assessment, student growth, and teacher of record all intersect in regards to children receiving special education services and their teachers. Many districts continue to wrestle with questions regarding appropriate assessment tools that will nevertheless promote growth, while determining teacher of record in what may be a more fluid teaching environment. This is further compounded by state initiatives, such as standards-based education, that will significantly impact assessment measures and outcomes for children enrolled special education.

> *Superintendent E: At the elementary level it's a little bit easier because you have a set class, but you still have situations between special education and regular classroom teachers. So we look at that 80 percent attendance and instructional*

level. We had some questions around team teaching because we do that between special education and regular classrooms.

Superintendent F: [Assessing growth in special education students] is one of the specifics that we need to come back to. There's also, concurrently, our work in proficiency-based, and how IEPs now need to be written in proficiency-based goals -- which is categorically different than probably what they had been before.

Several superintendents reported on how co-teaching students in special education, combined with a more inclusive curriculum, served to address PE/PG needs while simultaneously enhancing the educational experience for these students. In particular, they felt that co-teaching resulted in students in special education receiving a stronger curriculum that is more aligned with standards, leading to improved academic outcomes for these students.

Superintendent A: We have gone to co-teaching. So our special education teacher is co-teaching with the regular education teacher. We find for 85% of our special education population, this is working. The other 15% who are more severe, it's a little tougher and we may have to do more pullout. But we found that when we aligned our special education curriculum with say, our biology curriculum, [students in special education] weren't even touching what the other kids were hitting in biology. So, in the last 3 years now, we've been co-teaching, with special education and regular teachers together – and for 85% of the kids [in special education] it's working…

For the PE/PG student assessments, co-teaching may also lead to a significant portion of students in special education being able to have their academic growth measured using the same (or parallel) instruments as their peers who are not enrolled in special education.

Superintendent D: We like an inclusion model, and so special education teachers and regular content teachers are going to share the students. They'll both be teachers of record. We can't use the IEP goals, so whenever appropriate we're going to mirror the expectations for all students. There certainly are some self-contained students that don't do the same assessments that all students do. But we're going to have to look at what assessments we will use for the self-contained…. If they're not doing the grade level assessment [there exists a complimentary special education version for some instruments], so we try to get as many of the students as possible to take the grade level assessments. But if their IEP is such that they can't handle the grade level assessment, then they may have to go to the alternative assessment...

Superintendent A: Having these students exposed to that regular education instruction hopefully allows them to take the assessment. [Historically], our

special education population get extended time, they get a monitor. And if you didn't know it, they didn't know it. So extended time didn't matter… because they weren't exposed to the curriculum. By going to co-teaching we found this has helped us a little bit more with our special education students [who are now exposed to more of the curriculum].

Co-teaching though requires considerable work by the teachers and changes in how both educators see their role working with students in special education; but ultimately, superintendents reported that the result was positive for students and teachers alike.

Superintendent A: All in all, it took a while for the two teachers to see their roles [as] teachers. They're co-teachers. The special education teacher was not a glorified ed tech. They're a teacher…. A special education teacher who has a strong background in science may not have the strongest background in English. But if you can co-teach, [special education students] are going to get the strong background [from teachers in other disciplines] because you're going to be exposed to those teachers [more].

FACTORS IMPACTING STUDENT GROWTH AND TEACHER PE/PG SCORES

Not surprisingly, superintendents noted that student growth will be impacted by many personal, familial, and community factors that operate outside of the classroom. Family educational patterns and a history of higher educational attainment will impact student aspiration and motivation, which impact academic testing. Family and community values and work experiences will also impact the interest and goal setting for some students – are they interested in employment that involves a four-year degree, or local jobs that may require other skills not covered by standardized assessments? Failure to recognize these issues when examining student assessment and growth data can lead to misinterpretations of a teacher's performance relative to other teachers across the state.

Superintendent : [Years ago] we had 40% of our juniors and seniors taking the SAT before it became mandatory -- Bangor had 84% of their juniors taking the SAT. Then I started looking at aspirations, and a lot of our students just want to get through school and go work in the forestry sector… They're high-paying, hard-working jobs…so I think the aspirations of saying, "I'm going to a four year college and need to do really well on the SAT" [doesn't resonate] the same as someone in the Bangor area.. Then [with the mandated SAT] they went to 100% and they're looking at 38 to 40% of our people meeting or exceeding. We've tried to improve on that, and we've brought it up to the low 40s, but even that was a lot of work.

Superintendent D: We also have a situation where we have the highest mobility rate in the state. I have, for example, one teacher this year that works in my highest poverty school, 97% free and reduced, she had 18 children at the beginning of the school year, and at this point in the school year she has four of the original 18 that started. So you do a one-time high-stakes test {laughter} that's not fair. Her N size is so small, and two kids could have a bad day and it looks like 50%. Somehow you've got to understand that.

OBSERVATION AND OTHER STUDENT PE/PG DATA

While the previous sections have focused on student assessment data and measuring student growth, this section addresses other key data typically included in PE/PG systems. First, classroom observation data and observational systems used by these districts are examined, followed by a discussion of student surveys of the classroom environment. In some cases, these different types of data may provide conflicting or contradictory impressions regarding teacher effectiveness, and so superintendent perspectives regarding such discrepancies and how they are addressed in their PE/PG system are then examined.

OBSERVATIONS

Not unexpectedly, observations of teachers "in action" in their classroom were uniformly seen as vital components to assessing the quality teaching. Superintendents in these case studies reported using a variety of different classroom observational tools, based on different standards or models of teaching. Regardless of their selection, superintendents reported satisfaction with the observational system used by their district. All observational systems were computer/web-based, with a range of enhanced data collection tools to help observers make reliable, accurate assessments. These systems also included reporting tools to assist teachers and supervisors interpret the results and identify skill-areas in which a teacher may benefit from further attention and training. In some systems, these recommendations may be tied to specific suggestions and examples for the teacher to consider in future practice.

> *Superintendent G: We are using RANDA, which is a firm out of Tennessee. We're in our 3rd year with them…. All administrators have iPads and can record the observation on the iPad. It can be uploaded to the website -- it's a website-based thing – and they access it from a laptop or a desktop. It's something that teachers can access [and] communicate with their administrator. All the observation notes and anything the teacher wants to add for evidence are all in the system, as well as the ultimate summative rating.*
>
> *Superintendent F: So, with the Marzano model, because we chose to go that way, there is software called iObservation that's used to keep track of observation data.*
>
> *Superintendent G: It's the ability for teachers to get quick feedback, to be able to respond to that feedback in writing if they want, the whole timeliness of post-*

observation conference -- that's all taken care of in RANDA. Additional evidence the teacher may want considered can still [be included in] the face-to-face meeting, but I think it logistically has allowed us to do this. I'm not sure we could have done this just using paper.

Superintendent E: And what's great about that new updated Danielson version, I've been writing up evaluations the last couple of nights. The rubrics are so detailed that it makes it very clear. So I gave a teacher a 2 in the questioning area, but to me, when we have the post-observation it's going to be very easy for me to say, "You know, if you can provide me more information about this, I'm happy to up that to a 3. But I just don't remember seeing anything."… It gives [teachers] examples and everything, so I don't feel like teachers are debating their scores. The rubrics are very detailed. You can read very clearly the differences between a 1, 2, 3 and 4, and then under that there are two sets of further explanations. There are examples and critical attributes of those areas. It's really a great reference book for a teacher to look at... If I want to go from a 1 to a 3, what are the critical attributes that I have to have in my instruction to improve that area? And then there are some sample questions or strategies that [I] can use in [my] instruction. So it's really good.

Observations are generally conducted by administrators or supervisors, including principals, assistant principals, department heads and directors, or others depending upon the grade level and specific organizational structure of a school.

Superintendent G: The observers are an administrator. So it could be a principal, assistant principal, special education supervisor, English Language Learner director… That group would probably be doing 95% of the observations.

Superintendent F: At this point it's primarily supervisors. I use that term because in most places that's going to be the building principal, or where there is an assistant principal, the assistant principal. But I supervise a couple of teacher leaders, so in some cases it's me. There are also some of our special education coordinators that supervise teachers, and so it's them. At the high school, department heads are responsible for evaluation of their department teachers. So where there has been a probationary teacher in one of their departments, they have done the observations as well.

Superintendent D: Our observers in grades pre-K to grade 8 are administrators. In [grades] 9-12 would be a combination of administrators and department heads. And the department heads all have had to take the supervision and evaluation coursework at the graduate level.

Superintendent E: Administrators do the observations… -- building principals, assistant principals, special education directors. Administrators have been doing

observations and summatives all along, so we're keeping a system where that doesn't increase. I think the increase has really come with us getting our goal-setting more rigorous. And the student data piece will end up being more time-intensive, and [establishing] the teacher of record.

Some districts also include peer observations conducted by other teachers. These may be formal or informal, and depending upon the district they may be strictly for a teacher's own edification and not made available to administrators as part of PE/PG evaluations. Peer observations are seen as providing an additional independent perspective on a teacher's performance in the classroom, but also are seen as a tool for encouraging discussion, collaboration, and idea-sharing among teachers. For some superintendents, the most valuable contribution of peer observations was in facilitating this type of supportive learning community within a school.

Superintendent F: [For peer observations] we're using the same tool and it's the same standards that everybody's looking for and giving feedback on. The only difference is that when it's done by a peer, that information is not visible to the principal in the software. So if you and I are both teachers and I go and observe you, obviously you will be able to see what I had to say in the software, but our principal would not be able to see it. So it's just between the two colleagues and it doesn't rise up to the principal. It wouldn't become part of the evaluation.

Superintendent F: This has happened to a lesser degree, just because of the logistics of substitutes and releasing and so forth, but we have had a number of the teachers who've been to the trainings with us this year, [and they] have had opportunities to do the required peer observation/peer review piece as well. So we have had teachers go into other people's classrooms, using the iObservation software and the Marzano model to give their peers feedback.

Superintendent D: We have allowed peer observations as an option for those that are either effective, on, distinguished and on continuing contract.

While classroom observations provide valuable impressions of the instructional style and learning environment, establishing reliability for these observations can be a considerable challenge. To be valid, formal observational systems require training and reliability-checks in order to ensure that different classrooms evaluated by different observers are nevertheless being evaluated in the same manner using the same criteria and scaling. The potential that even well-intentioned observers may improperly rate classrooms and negatively impact PE/PG evaluations is reportedly a common concern. Unfortunately, the costs for implementing an observational

system and training observers to reliability can be significant. This can be offset in part through partnerships with other districts implementing a common system.

> *Superintendent G: Teachers are still and probably will forever be concerned about reliability, and we've devoted a fair amount of our administrative team time in the past couple of years to training -- whether it [involves] reviewing videos and scoring them collectively, or just dealing with different aspects of the evaluation process.*
>
> *Superintendent F: We have actually joined efforts with probably half a dozen other districts that have also selected the Marzano model to share the cost of training. And so for our administrators as well as a good number of teachers that were interested in coming along with us, we have had four and a half days of training this year on observation using the Marzano model. So, that's been really, really valuable.*
>
> *Superintendent: RANDA is roughly $30,000 a year. Now, by Maine standards, [we are] a large district. We've got over 400 teachers. And, you know, if we were a district that had under 100 teachers, as many districts in Maine do, it might be quite different. But we didn't see how we could possibly maintain this system going forward without using something like RANDA. I know some districts have developed their own database. RANDA has been responsive and again, I think consistency is an important element to successful administration of anything. And so our teachers and administrators are familiar it now and I'd be reticent to change it at this juncture.*

Given the costs involved, superintendents uniformly reported that support from the state for observations would be particularly valuable. As noted in previous comments, costs for high-quality systems may be acceptable for larger districts, but prohibitive to smaller ones, placing those districts at a relative disadvantage. Other related support, such as state-sponsored professional development or regional training, or state assistance coordinating larger collaboratives using a common system would be additional ways the State could provide help.

> *Superintendent F: [Additional] money should be earmarked for supporting this. That would be great. I know the education committee doesn't have final say on that, but they could talk to their appropriation colleagues about the fact that we would love to have not only money for a proficiency-based system, but also the PE/PG system that were part of the governor's budget. We would certainly appreciate that….[Funding would be helpful with training and professional development], software that we use to keep track of all that, and substitute*

teachers when staff are out at trainings as well as when peers observe each other. Those are three major costs that we incur for this.

Superintendent E: We could always use funding around the teacher evaluation system. Even if we were using it for the peer observation piece and to learn from others. There there is a cost. We've had to release teachers to meet on the committee.

STUDENT SURVEYS

Some districts also incorporate student surveys into their PE/PG system. Student surveys provide an additional unique perspective on the classroom environment and instructional practice. For the purpose of assessing teacher effectiveness, surveys should focus on the teaching environment and pedagogical style of the classroom, rather than simply address student learning or satisfaction with a curriculum. Depending upon the district, student surveys may be widely implemented or used on a limited scale in response to specific concerns, such as contradictory PE/PG data.

Superintendent D: We have student surveys, but we've created our own student surveys. We did look at the Seven C's survey. But for the amount of money we didn't think it did all that it was supposed to do, so we created our own surveys. They are optional unless an observer feels that [a student survey] should be used given where the teacher lands for certain core propositions. So, if they're not committed to students and their learning, or if they are not knowledgeable of their content and know how to teach it to their students, and we feel there's a problem, then the observer could say that you need to utilize student survey to get data from your students.

Not surprisingly, student survey data regarding the classroom experience introduces new issues. Specifically, teachers may be uncomfortable with the idea that their teaching effectiveness is being "evaluated" by their own students, and concerned that student ratings may be swayed positively or negatively based on other factors, such as grading, homework expectations, etc. Superintendents reported that these concerns can be ameliorated – to at least some degree – by using appropriate instruments and clear communication regarding the nature and purpose of the student surveys. While it may be painful when student reports differ from teachers' own self-perceptions, the resulting process of reflection and change was seen as beneficial to both the students and the teacher.

Superintendent B: Some of the teacher feedback was based on a misunderstanding of statistical analysis, and a misunderstanding of how data works. But it affected the teachers' perceptions of the results. Another level of misunderstanding was a fairly basic one, and one that's tied to language. Teachers would often be heard saying things like "well the students are evaluating us", when in fact that's not what's happening. Students are reporting their perceptions of the learning environment, which does include the teacher, but, [other areas, such as] classroom management.

Superintendent B: Initial reactions to the results from their students were almost traumatic for some teachers whose self-perception did not agree with the reported student perceptions. And there was a significant period of time where some teachers were grappling with this disconnect... That, in and of itself, we see as a positive thing. The teacher began to reflect more deeply on why that was occurring, what was going on, hopefully coming to the realization that the student's perception of the classroom was the student's reality of the classroom, regardless of the objective reality. It didn't really matter if most of the students know the rules and follow them if the student's perception sincerely is that students don't know the rules and don't follow them…That was a difficult thing and is still a work in progress for people to come to grips with. It's not always about the objective reality. It's about what the student's experience and perception of the reality is.

WHEN DATA DON'T AGREE

Not surprisingly, observation data, student growth data, student surveys, and other information contained within the PE/PG system may at times appear to be inconsistent or contradictory. Often, it may simply be an aberration that becomes clear once one reviews more long-term data and trends for that teacher. Other times, it may reflect more subtle and complicated issues. For example, an excellent teacher may correctly appear very strong based on supervisor and peer observations, and yet if she teaches in a highly mobile district – where mobility can negatively impact student performance (see MEPRI report ZZZZ) – her student growth data may appear problematic. Similarly, the actual content in a class may not align with the instruments being used to assess student growth. Alternatively, observations may identify numerous concerns for a teacher, and yet if he teaches in an academically strong school or community, he may have a large proportion of students identified as "proficient" simply because he is "riding along on good demographics".

Superintendent G: I'm not sure inconsistency is the right word. I guess I would say this. We have excellent teachers who don't necessarily have stellar results in growth, and vice versa. And part of it is that this is a complex [issue]. I've been involved with performance-based evaluation and compensation systems [in other types of work settings] and education is complex. For example, I remember a teacher who was in one of our most needy schools, with 18 kids in her classroom and her top four kids moved in March… Mobility is a big problem. We have some classes where 35% or 40% mobility is not unusual. In a perfect world, a student would be assessed the day they leave a school and the day they start the next school. We're obviously not going to be able to do that. So you end up, trying to pro-rate and it's not a perfect system. I will say this though, that the teachers that have come out of our system as either being ineffective or developing, or on the other end [are identified as being] distinguished have a preponderance of the evidence across the board that has led to a reasonable conclusion that the assessment is correct.

Superintendent A: We look at the test, what is the assessment information? What are they testing kids on? And are we actually teaching that in the course that we say we're teaching it in? That's a huge piece of alignment from one grade to the other.

Superintendent E: I'm not saying that it doesn't ever happen. One principal at one school called and said that the teachers are shocked when they looked at their own data… and we feel the teachers are really strong. So I think it can happen. But I don't think [it will happen] over time and that's what I keep stressing -- we want to do data because any one year's data point doesn't mean lot. I'm looking at the trend over time. And usually trends over time don't lie.

In particular, superintendents expressed concern that the alignment of assessments with coursework will require particular attention over the coming years. In some instances, course material may not align with an assessment simply because a teacher, for one reason or another, chooses to teach material outside of the official, well-aligned curriculum. However, the transition to standards-based education is seen as potentially more fundamentally changing the content and timing of educational material for all students. With every change, districts will need to monitor their curriculum and assessment instruments in order to verify that what is being measured actually aligns with what is intended to be covered in the classroom.

Superintendent A: And well, [every district has a few] teachers who teach what they like to teach. And it may not be part of the Common Core. It may not be part of what we need. And we may need to have [the teacher] change that. So,

it's a definite issue that hopefully this assessment, this teacher effectiveness will help us with.

Superintendent D: As we move to proficiency-based learning, or standards-based [learning]…the NWEA may not be the best measurement of that. And so I think, increasingly, we may be measuring something that doesn't necessarily match up with what we are expecting our teachers to do in the classroom.

Superintendent G: I think, increasingly, we may be measuring something that doesn't necessarily match up with what we are expecting our teachers to do in the classroom. If the MEA is truly aligned with the Maine Learning Results, then that [would be ideal]. With all the angst in regards to assessment, time will tell, but I think as we move more to proficiency-based learning and the use of common assessments by our teachers, we're going to be in a better position to make that determination with our local assessment system. Until the common assessments are developed and in use at all grade levels, it's really tough to draw any conclusions from what's happening in a class or in a district. But we are increasingly moving that way. And if a district that does indeed have a strong common assessment system [they] arguably would have less need for any other outside assessment.

DIFFERENCES ARE EXPECTED AND DESIRED

Finally, when the information within the PE/PG system is ultimately analyzed, the results should ultimately show differences between teachers. To be accurate and useful, a PE/PG system must differentiate between teachers with some teachers recognized as higher performing or more effective than others. In theory, when a single assessment instrument is examined across an entire state, it is possible for every teacher in a given school to be rated above the ***state*** average – but any system which shows no differences between teachers is of limited value.

Superintendent B: Since we've been increasing our attention in terms of implementing the new professional practice standards and including the other measures, we've seen an increased differentiation of the assessed performance. Even on the professional practice side. Initially, we had a pattern of results that was strikingly similar to the traditional pattern of everything being right-shifted on the bell curve. [Now] what we're seeing is greater differentiation…. We have much less 'everyone falling into the same category'… Now, that doesn't guarantee accuracy, but at least it appears we're measuring some difference now.

Other superintendents noted similar differentiation when looking at classroom level data. One district creates visual displays of their results. A color-coded bar shows the percent of students

at different proficiency levels in a classroom, with green indicating the proportion of students who are proficient in that topic.

> *Superintendent C: At times you'd see the green band [of proficient students] shrink over the course of the year. So say [the students] had a dynamite teacher last year, there's going to be some decrease in their proficiency over the summer. But if the green band is [steadily shrinking], it puts you on the principal's radar. It should.*

While some teachers understandably find this type of evidence-based differentiation stressful, superintendents also reported that other teachers were positive about different levels of performance being recognized. Particularly hard-working and high-performing teachers may be frustrated by a system that simply places all teachers into the same category. Furthermore, it is difficult to target and address the need for additional training and support if the evaluation system fails to flag those teachers in need of such support.

> *Superintendent B: I think there's been a mixed reaction, and it depends on where you stand. My belief is that teachers historically have felt uncomfortable -- especially teachers who were more effective and more successful with their student population -- with all teachers being categorized as above-average. Many teachers are feeling more positive about a system that actually recognizes differences in performance, and recognizes that some teachers are more highly performing than others. There are some teachers who I believe have been uncomfortable with that identification... We've had a couple of teachers in the last two years who have had individualized support and improvement plans that have arisen from the system. And in both of those cases, at post-intervention there has been a demonstrated increase in the teacher's effectiveness as measured by the student perception survey, student learning outcomes, and by professional practice assessment.*

Superintendent *D: We should **not** make the focus on the teachers that are not doing their jobs. We have to make sure the system is robust enough that we're helping **all** teachers grow…. a robust professional development system that's built on teachers continuing to think about their craft and how to grow to the next level. [We want them] to have that excitement about finding ways to have a tremendous impact on not only their students, but students across the district.*

MAKING IT WORK: SUPERINTENDENT SUGGESTIONS

While the focus of this report is student assessment and growth data in PE/PG systems, superintendents also noted several common strategies they felt were particularly valuable when developing and implementing their systems. This final section summarizes their observations in order to help inform others who may either be at an earlier stage of building a PE/PG system, or redesigning their system in the future.

DON'T DELAY AND STAY THE COURSE

These superintendents as a whole felt that their districts were in a relatively strong position given they started early and/or worked aggressively to meet target dates. Superintendents also recognized that starting early meant that they had to make subsequent revisions to their plans based on changes implemented by the state of federal government – some of which were made in order to accommodate districts that were late in starting.

> *Superintendent B: We actually began this work around student learning and growth measures in our performance pay system. In that system, we use a much broader set of student learning and growth measures, many of which are not permitted within the evaluation system, including such things as on-time graduation, student attendance, et cetera. [These] are not consistent with the rule in regards to measuring student learning outcomes, but are broader outcomes that are desirable for the school as well. We set up targets that were at the individual level, being… students which X teacher had within their instructional cohort. [We also established] targets at the team level, which included targets for the broader base of students that are instructed in a given content area by a team, such as the secondary mathematics department. And*

[we established] school-level targets and district-level targets. So we had actually four levels of targets. Each teacher had a set of goals that included all four levels at some point within their goal set, with no less than 50% of those and no more than 75% being individual measures. So we had done a lot of work early on around a broader idea of student learning and growth measures…[and] worked through those things collaboratively right from the start. We made mistakes, and we had a process that during a cycle we live with the mistakes we make, but at the end of the cycle we shall be brutally self-reflective in terms of tearing apart what we did in the last cycle to find ways to improve it, and analyze it and abandon things that we [previously] thought were good ideas at the time.

Superintendent F: Although there have been some good tweaks, what has been extremely frustrating is the number of changes in the rule as we've gone through the process. So I guess, one thing that's good is that we started out early. One thing that's bad is that we started out early. And so although I understand why some of the changes were necessary, and some of it came from the Feds, and so forth, it's been pretty frustrating to have to go back and say okay where are we now that the changes have been made? …..Let's just kind of stick with it for a little bit so that we can get things in place, and know where we stand.

The result of is that districts may feel that time and resources were wasted designing and implementing changes that were unnecessary. This may discourage districts in the future, or create an unintended disincentive to readily adopt new policies in anticipation that these policies will go through multiple revisions or delays.

Superintendent A: This has been on the books…. Schools have gone out and are either at a point that we are at, or even better than us. They've [developed a system]. They're using it for a couple of years now with everybody. You know?... There are schools that haven't even begun to talk about this and it wouldn't be fair to the other schools to change the rules all the time. Because then it makes it look like, "why did we do all that work? And now the state says we don't need it." Stay the course.

Superintendent G: I've said this to the committee at different times and different points: consistency is critical. That includes consistency in state policy and law. There can be some small tweaks that may well be appropriate, but, we're seeing now, with proficiency-based learning and all the different laws or bills that it's just creating uncertainty. Districts or schools as institutions are slow to change. If there is no consistency of policy, we're going to fall short in our implementation.

DON'T RELY ON ONE SOURCE – MULTIPLE SOURCES OF INFORMATION

Superintendents also reported finding value in drawing from multiple sources of information for all components within their PE/PG system. In some cases, evidence of high (or low) quality teaching may reflect low-frequency events that occur in specific situations. As such, different tools may capture different, unique insights into a teacher's practice. Multiple sources thus lead to more reliable and valid summaries of teacher performance and effectiveness. It also serves to address teacher concerns regarding potential problems or biases with any single source.

> *Superintendent B: Anytime we measure with a single yardstick we don't have full confidence in its accuracy and reliability. That's a problem. So we want to use multiple yardsticks that have some level of confidence of doing the job of measuring, and consider all that data together over time to be able to feel a higher level of confidence in our assessment… we include both the supervisor's assessment of professional practice, but also a portfolio from the teacher of their own demonstration or evidence of professional practice standards, [as well as] student perception surveys, [and] student learning outcomes from more than one source.*

> *Superintendent D: We use data wisely – and not to get into a trap of thinking that you can look at a simple score of achievement and truly understand the quality of the teacher. You have to look at that data set from multiple perspectives before you can truly determine the effectiveness of a teacher… We are totally committed to triangulating the data, and not using one high-stakes assessment.*

Triangulating information across multiple measures fits well with a holistic perspective teacher effectiveness. This can reconcile possible conflicting pieces of information from different instruments collected in different ways. All information then needs to be integrated into an overall multi-dimensional summation of teacher effectiveness.

> *Superintendent D: We thought that it was better to look at a model that is a holistic type of a scoring, such as the way you score writing, in which you use a preponderance of the evidence. If you look at "committed to students and their learning", as a standard, you may not be able to observe that with every student in every lesson, even with walk-throughs. But you certainly would work with the team of the teacher and the observer to articulate the number of ways, and what types of evidence can we look at to show that you are committed to the students and their learning. You then could look at each of the standards.*

DON'T SLOW DOWN – MEET REGULARLY

Once the process of designing a PE/PG system is in place, the superintendents reported that to keep the process moving forward, regular meetings among committee members are vital. Communication and meetings with other districts are also valuable ways to learn strategies for addressing challenges that may emerge, as well as for leveraging ideas or tools that other districts have identified or developed. Not surprisingly, superintendents also credited success with having open, inclusive membership in the design process.

> *Superintendent E: It's been really helpful for us to meet at least every month with our group. If you don't meet with them on a regular basis, you forget where you're at and you lose that momentum….[Also] we belong to the Western Maine superintendents' group, and it's been helpful just for us to be able to get together and talk about these kinds of issues. You know, what are you doing related to this or to that? And how are you handling this or that? So that's been helpful.*

> *Superintendent B: Right from the beginning the majority of the individuals working on this were faculty members – the evaluation team consists of eight teachers and two administrators, so a little over 10% of the faculty. But the rule in terms of membership, was in essence a coalition of the willing. It's an open membership. The superintendent doesn't appoint people. The association doesn't appoint people. Anyone who is willing to commit to the process and commit to a collaborative approach to achieving the goal of establishing an equitable and valid basis… was welcome to join at any time, at the level of participation that they could commit to. That structure I think contributed a lot to the success, in terms of moving it forward.*

> *Superintendent B: Three years ago, we did a presentation at Maine School Awards fall conference… and someone from the audience asked a question similar to "why is this working?" And [the presenter's] response was: in the past, collaboration has meant we were invited in to help decorate the cake. This time we got to be part of picking out the ingredients.*

COMMUNICATE

As the PE/PG system design and implementation progresses, it is important that there be regular, clear, and consistent communication with district teachers and administrators. This is important in order to ensure transparency as well as to identify and correct any misconceptions that may arise regarding the process or goals. Ongoing communication, including information and

examples for various instruments, can also help quickly address questions that may arise and provide support and guidance to teachers or administrators as they prepare for the change.

> *Superintendent E: I think another challenge is the communication. We do things at the committee level. We need to communicate that out to administrators, to the staff. We need to keep the board informed, and we've all got to be consistent in what we're communicating and what we're doing. We've tried to do a good job of communicating the same message, and it's been helpful having teachers on those committees, because we've got representation from all the schools. So when I go to the high school and give a presentation and explain this… I've got a couple of teachers sitting there that have also heard how this is supposed to be explained and presented and used.*

> *Superintendent D: We just asked [faculty] "What part of the [PE/PG process] do you feel will be most helpful to you?", and I think the general [sense] is that they feel they will have more specific feedback--that that's what they like. [We also asked] what else do we need to do to help you better understand this system…, and overwhelmingly it's 'we need specific examples.' Although everybody has written SMART goals {i.e., Specific, Measurable, Action-Oriented, Realistic, Time-Bound} this year, they still want more examples of how to write SMART goals. They still want more specifics about the standards and how to collect evidence for those different standards. They want clarification of how, when it comes all the way down to each of the individual indicators under the core propositions, how do you get to preponderance of the evidence? I think back to when this state started scoring writing, those were the same types of questions. We needed anchor papers. We needed rubrics.*

AN ONGOING PROCESS

Superintendents reported that it was important that all teachers see PE/PG as a continually ongoing process, not just a "hoop to jump through" every few years. Having everyone engaged in some type of PE/PG activity each year was often seen as important. As goals are met or concerns addressed, new goals and targets should be established in order to maintain the momentum for positive change.

> *Superintendent E: We're making sure that every year there's some kind of formative feedback. In the past what's really happened is that teachers have had what I call 2 years off, and then in their 3rd year they have this "Oh, it's my year, right?" So we're trying to move from that model to every year you're going to have an observation [or] you'll have goal-setting -- how are you doing on that, [or] we'll look at the student data piece. So I think that's going to be the switch.*

Superintendent D: But something that we do, I {laugh} remember one of my new principals, he was all excited because he met the first annual SMART goal in reading. They met them at the half-year mark. So I allowed him to celebrate, and I agreed, and, you know, we were yahooing, and then I said, "So how have you adjusted the SMART goal?" And he said, "What do you mean?" And I said, "Well we don't rest {laugh}, so what's next?" [Their] goal that year had been that they would get—I think—95% of their kindergarteners reading at level C by the midyear point because level C is the end of year. And so he got 95% to level C and thought touchdown! I'm all done! Well, no no no! We've got a second half of the year. We need to do something here. So then what we did is set the SMART goal based on the first grade Fontis and Pennell benchmarks, so that they could keep going. And so it can be that you adjust a goal for even half a year.

HELPING TO IMPROVE TEACHING FOR ALL EDUCATORS

Superintendents uniformly reported that a central feature to a successful PE/PG system was that it not just be seen as a punitive tool used to discipline teachers. The goal is instead to help improve teaching, and through this improve student learning. As described in the opening quote for this section, a positive PE/PG system will be focused on building strengths and skills for all teachers, not just those who may be struggling. Superintendents acknowledged that in some cases the PE/PG process may ultimately lead to recognition that teaching – like any vocation – may not be the correct match for everyone who enters into the profession. But that should not be the goal of the process. That perception interferes with it being used to help promote better teaching in all educators.

Superintendent E: The premise of the evaluation system, and it's stated in there, is to improve teacher instruction and student achievement. So it's not there as a disciplinary tool or anything like that. It's to help and support.

Superintendent D: Unfortunately, out of the gate, all the negative communication around it has made even top teachers nervous. And it saddens me that that's the way it had to come out of the chute. [Because of this perception] we have to shrink the possibilities, build confidence for a while, and then allow it to grow. But when my best teachers are nervous, I'm like "what is this about?" {laugh} But it's because they hear comments and unfortunately both in the state and nationally, there's an attack on public education and the law and the system builds anxiety before you even roll out what it's about.

SUMMARY

Beyond validity, reliability, and alignment with the curriculum, superintendents noted several additional key features that were considered when selecting measures of student academic growth. These included:

- The time and scheduling of the assessment: it should provide maximum information, but require minimal disruption to the normal classroom schedule and practice.
- Tools that allow teachers to identify individual student strengths, weaknesses, and learning-gaps. This could be used by teachers for individualized learning.
- Potential for use as a formative assessment.
- Access to tools that administrators could use to inform policy and practice.
- A design that avoids even the potential appearance of possible data manipulation.

Beyond mathematics and reading/writing, superintendents indicated that their districts continued to seek standardized measures in other content areas. For example, some districts are using *Fit Stats* – a physical performance assessment already used by many schools in Maine—as a measure of growth in physical education, while others are addressing growth in the performance arts through change in portfolios or common performances over time. To identify measures, some districts have drawn upon instruments developed and used by multiple different sources and state PE/PG systems. Superintendents are also leveraging other existing or upcoming assessment efforts, such as RTI and proficiency-based education, as a way to address PE/PG assessment needs. Partnerships with other districts, as well as guidance from the Department of Education, can help identify such solutions. This can also help districts avoid "recreating the wheel", as well as potentially help offset the cost of researching and implementing solutions.

Superintendents expressed a number of concerns with incorporating the MEA / Smarter Balanced assessment into their PE/PG systems. Foremost was the fundamental question of whether the Smarter Balanced assessment was going to be used beyond its first year. Districts were uncomfortable shaping the student-growth portion of their PE/PG system around a tool that may only be in place once. The degree to which superintendents felt that key information regarding Smarter Balanced results continued to be unclear was an additional major concern.

These included alignment with curriculum and future policy, such as proficiency-based education, as well as the type, extent, and format in which the results would be provided to educators. Also, districts that had institutionalized a formal vetting process for reviewing and selecting instruments were hesitant to adopt a new tool without applying the same standards and review process to a new MEA measure. While Maine has since decided to discontinue using Smarter Balanced, these concerns may prove valuable when selecting a new measure.

Superintendents observed that the alignment of assessments with coursework will require particular attention over the coming years. The transition to standards-based education may lead to significant changes in the content and timing of some material. Districts will need to monitor their curriculum and assessment instruments in order to verify that what is being measured actually aligns with what is intended to be covered in the classroom.

Some districts assessed growth over different time scales based on the specific course, with growth over a single year applied to courses that are covered on a regular, steady basis (e.g., mathematics), and growth over multiple years used for courses that have more limited instructional time (e.g., performance arts). Districts also varied in how they addressed summer learning loss. Spring-to-spring assessments include any loss that occurs over the summer: The greater the summer learning loss, the greater the improvement needed to "break even" with the previous end-of-year spring assessment. However, as noted by one superintendent, ignoring summer learning loss may overstate how much ***true*** growth is occurring over multiple years and unintentionally lead to schools not exploring solutions to this issue.

Districts included in these case studies generally weigh student growth as 20% or more of a teacher's PE/PG score, or are building to 20% over the next few years. Individual student growth was aggregated at different levels in different districts. For example, a teacher's student growth component of the PE/PG score might be based on all the students in her class, all the students in her academic program, all the students in her grade-level, and/or all the students in her school. In some districts, the growth component for a teacher's PE/PG score may be based only on the classroom-level aggregation, in others it may be based on all four aggregation levels. Alternatively, if she does not work with a specific class, it may be weighted to focus on student growth in her program area or grade.

One of the challenges with incorporating student performance data into a PE/PG system is identifying an official teacher of record. Superintendents reported that it was important to include a degree of flexibility in assigning teacher of record in order to address unique situations that may arise, while recognizing the need to carefully evaluate and monitor such exclusions in order maintain the validity of the entire system and avoid "cherry-picking" student scores. A deeper concern was that too much attention on the teacher of record may lead educators to focus solely on children who they perceive as "their" students, at the expense of providing support and assistance to other students around them. This was particularly true in regards to students in special education, where several districts relied heavily on co-teaching.

Districts use a variety of classroom observation tools based on different standards or models of teaching. Superintendents were uniformly satisfied with whatever system their district used. Observational systems were computer/web-based, with features to help observers make reliable, accurate assessments. Systems also included reporting tools to help educators interpret the results and identify skill-areas for future professional development. Some districts also include peer observations, which depending upon the district, may not available to administrators as part of PE/PG evaluations. Peer observations were seen as a useful way of promoting discussion and idea-sharing among teachers.

The significant costs for implementing an observational system and training observers can be offset in part through partnerships with other districts, although superintendents uniformly reported that support from the state would be valuable. Other possible state-level support, such as state-sponsored professional development or regional training, or state assistance coordinating larger collaboratives was also seen as potentially helpful.

Student surveys were part of some state systems, and depending upon the district either widely implemented or used on a limited scale in response to specific concerns.

Ultimately, while some teachers were uncomfortable with PE/PG systems identifying teachers as performing at different levels of effectiveness, superintendents also reported that other teachers were positive about different levels of performance being recognized. Particularly hard-working and high-performing teachers may be frustrated by a system that simply places all teachers into the same category.