

# **Implementing the New Educator Evaluation System in Massachusetts**



**A Coalition of Twenty-Five Massachusetts Educational Stakeholder Organizations**

## **EXECUTIVE SUMMARY**

The new Massachusetts Educator Evaluation System lays a sound foundation for transforming the educator evaluation process. However, while structures and regulation can facilitate change, it is the deliberate behavior and actions of people that ensure that such changes produce improvement; it's all about implementation and the values and beliefs people bring to the implementation.

This paper analyzes the positive components in the new evaluation system, identifies possible pitfalls that may occur when implementing the new structures, and outlines the necessary supports that must be in place to ensure the system is robust and used effectively to bring about substantial change in the profession.

### **I. WHAT'S GOOD ABOUT THE NEW REGULATIONS?**

The new regulations allow shifts in district and school culture that result in a focus on student learning and educator growth. That shift enhances the education profession. Everyone is now required to pay attention to certain kinds of expertise that really are legitimate parts of the skill set of a high-functioning teacher, but which were not included before, or even explicitly excluded, in many evaluation systems. Moreover, an emphasis on educator responsibility for student learning undergirds the new system, thus codifying a tenet of effective teaching.

#### **Positive #1: New Categories**

##### **Family and Community Relations**

- Reaching out for healthy productive relations with parents and community that benefit student learning.

##### **Professional Culture**

- Being a contributing colleague and a team player; that is the center of what is meant by the Professional Culture standard.

##### **Cultural Proficiency**

- Becoming culturally proficient and showing it in your instruction and relations with students. That means going beyond holidays and ethnic cooking to planning culturally relevant lessons, using literature from the students' culture, having the culture of the students show up in the artifacts of the room, finding out about the history and values of the students' families.

##### **Analyzing Data**

- Knowing how to collect and interpret data and use it to adjust instruction, including applying new strategies to reach students who didn't "get it" yesterday.

All the above are important areas of performance for successful teaching and are included in our new system. They now become part of our common definition of what a highly effective educator does.

### **Positive #2: Student Test Scores Not a Numbers Game.**

In the Massachusetts system, student results are included in a responsible way. We have not been drawn into the numbers game like other states where an arbitrary percentage of one's evaluation (50%...33%...60%) is pegged to showing an increase for students on the state standardized test.

We do have to eventually rate a teacher's effect on student learning as low, moderate or high, but with no weighted percentage attached, and we will use multiple measures to show the teacher effect. Should a teacher receive a rating of low on student learning but an overall ranking of proficient based on goals and standards, a conversation between teacher and evaluators must dig into this discrepancy. The action required will be that the teacher receives a one year self-directed plan instead of a two-year plan.

In addition, "the Educator Plan is expected to include one or more goals directly related to examining elements of practice that may be contributing to low impact. The evaluator's supervisor must review the performance rating with the evaluator when a notable discrepancy occurs. When there are significant discrepancies between impact ratings and summative performance ratings, the evaluator's supervisor may note these discrepancies as a factor in the evaluator's evaluation." These are responsible guidelines without being heavy handed.

Given the way the timelines are laid out and the need to have two years' worth of data before making a judgment about a teacher's impact on student learning, it is not until 2016 that we will be giving most teachers ratings on their impact on student learning. This is wise planning, because it gives us time to unpack the phrase "multiple measures of student learning, growth, and achievement" and generate physical samples to share with one another of different kinds of evidence that would count. An important caution is the implementation of District Determined Measures (DDMs). Districts will require clear DESE guidance and the time to collaborate in the identification and development of DDMs. Additionally, these assessments will need scoring reliability established across the district.

### **Positive #3: Unannounced Observations**

Some important obstacles have also been removed:

- Unannounced observations are now allowed everywhere. Dog-and-pony show lessons for evaluators finally recede into history. This provides a more authentic view of the teaching and learning that is happening in every classroom and potentially allows evaluators to gather evidence of educator growth.
- Dismissal for teaching that is unsatisfactory for a person who has received intensive assistance is expedited to one year. But this could have uneven application in some places. Depending on what was negotiated, some districts could immediately tell a veteran teacher that s/he needs improvement and give him/her as little as 30 days

to improve. Then they could label the teacher as unsatisfactory after as little as 30 days, meaning that in some places one could potentially evaluate out a veteran teacher in as little as 60 calendar days without the time for intensive assistance that is intended. Prior to this, however, we would expect that a person be on a Directed Growth Plan, because s/he had already been identified as needing improvement, and had received appropriate assistance and sufficient time to use it.

#### **Positive #4: Teachers Actively Involved**

Teachers are now actively involved, no longer passive recipients, in the evaluation process.

- Self-evaluation is now a substantive part of the process for teachers at all stages of their careers.
- Every teacher has to create a growth plan that makes him/her accountable for one goal related to improving a specific professional practice and one goal related to student learning.
- Teachers have to gather and submit their own evidence of student growth. This makes us all more results-focused without hamstringing us on how we will report that.
- Individual teachers are encouraged to set Team goals for which they are jointly accountable with members of common planning time teams or other appropriate teams. This feature encourages collaboration.

School and district leaders are also being evaluated through this process and the rubrics are aligned to promote common focus throughout the school district. The intent is to increase the level of focus on good teaching.

All of these features in the new system create an opportunity, if properly handled locally, to make educator evaluation a true growth-oriented process and not a judging and ranking exercise. They also create the need for efficient systems, robust training of all staff, and the provision of time at the local level to ensure that the benefits of these new components are not diminished by confusing and cumbersome application.

## **II. IDENTIFYING POSSIBLE PITFALLS**

1. Those implementing the new system must be aware of and resist the tendency to over-focus on compliance with every element rather than identifying and using the most productive aspects to focus on in the new system for the improvement of teaching and learning.
2. The public and those in the profession need to understand the purpose of the phrase “needs improvement.” It must become safe, normative, and expected in a culture of constant improvement, no matter how experienced and competent a person may already be. The public, policymakers, press, and members of the profession also need to understand the phrase “proficient.” Proficient is the desired outcome for the vast majority of educators. It means you have worked hard and achieved the desired results. Only a very small percentage of educators will truly be exemplars for others

and merit a rating of “exemplary.” Those receiving a rating of proficient should celebrate their success for a job well done.

3. The focus on the procedures, using the rubrics as a checklist, and electronic devices may lead to continuing superficial evaluations rather than creating a more substantive dialog that impacts student learning.
4. Hyper-focusing on deadlines rather than providing a flexible schedule for implementation that recognizes the variance between district readiness and capacity for implementation will short-change the process. This work is a marathon—not a sprint.
5. Study, training, and practice by evaluators to develop accuracy and significance in observations and feedback, in addition to inter-rater reliability, are necessary for fundamental change to occur.
6. The bar must be raised for high-expertise teaching and support for developing it must be in place to maximize the effectiveness of the new system.

None of these are trivial tasks; they call, in fact, for significant culture change and skill development for teachers and leaders in the profession, which has ramifications beyond educator evaluation.

The new regulations, rubrics and support materials aim to make educator evaluation a better process with a positive impact on teaching expertise and thus the experience of children. We applaud that goal and the good work of many players that has gone into the changes. But expecting regulations and structures to improve teaching is like expecting the Civil Rights Laws of the 60s to eliminate racism. No way. The Civil Rights laws were a significant push in the right direction, and they eliminated a number of structural obstacles to the prevention of discrimination.

Our new regulations are also a big push in the right direction, and they eliminate some obstacles to effective evaluation. But the work of making evaluation more meaningful and improving teaching practice across the board is up to us in the districts: teachers, teacher leaders, union leaders, building administrators, and central office leaders. No small task, and one that will easily take 5 years.

### **III. OVERCOMING PITFALLS THROUGH THE CREATION OF SUPPORTS**

#### **1. Getting hooked on compliance**

There are many pieces to implement before one could claim “We are complete!” for the new educator evaluation system. Districts need to ask what parts of it will be most productive for their students and focus on them first, rather than try to implement everything at one time.

Getting all educators to self-evaluate thoroughly and with understanding of the rubric could consume a year all by itself and will only happen if the district has worked to build a safe, trusting environment where teachers can be honest about their practice. But look at how many other parts there are to this system: getting evaluators to be accurate and reliable; learning not only how to write SMART goals, but how to make rigorous plans of action to actualize them; learning how to collect and record evidence from multiple sources that bear on an educator's performance; doing short classroom visits and gathering data that is significant; learning how to give useful feedback to teachers who are at different levels of professional maturity (Differentiated Conferencing); deciding what kinds of evidence besides standardized tests to use to judge a teacher's contribution to student growth; building a portfolio of examples. So careful choices tuned to local needs that prioritize the highest leverage aspects of the system should be made and used to guide the pace and sequence of implementation tasks. This requires a growth mindset. We will prioritize, learn from the process and get better over time.

Another pitfall of the over-compliance mentality would be to rush ahead with procedures and training without communicating clearly to all staff members what the district's hopes and aspirations were for outcomes of the new system. These hopes should be clear, desirable, and motivating to all staff, like "develop our ability for authentic, constant learning from one another and non-defensive self-examination of our own practice in relation to student results."

We offer the above only as an example. The point is: spend time creating the vision of the positive outcomes we want the new system to have for our students, for our educators, and for the spirit and cohesion of our district. Ask: What initiatives that we are already committed to can be augmented by how we implement the new evaluation system? And answer these questions clearly and often in communication to staff.

## **2. Failure to address the cultural significance of the meaning of "needs improvement"**

Strong adult professional cultures are characterized by constant learning and improvement of professional practice. This requires an acknowledgment that no matter how experienced or competent one is, there is always more to learn. Improvement never ends. So we ask: is it not natural that in a complex profession, a vocation where the knowledge and skill-base to function at a high level is as large and deep as that for architecture, law, or engineering, that one would "need improvement" in at least a few elements? (As opposed to standards—remember that a teacher must receive proficient in the first two standards to achieve an overall rating of proficient.)

Why then is it such an affront or cause for worry to many teachers when they are evaluated as "needs improvement" in certain areas? The answers are several and

will follow. But the point we will make is that this culture must be broken if the new evaluation system is to have the effect its structures promise. And all of us need to make it safe for educators to step up to the “needs improvement” counter with more humility and less fear.

One cause of this reluctance is the absence of norms of public teaching and continuous improvement amongst the adults in many of our schools and districts. Developing that norm of seeing one another practice could be the single largest liberator of openness and quality conversation among professionals. Working in isolation has become a habit of practice and observation is a rarity. Visits and conversations about practice are thus so rare that when an evaluator visits it is an “event” rather than a normative practice. Thus teachers get nervous, especially since the history of infrequent visits is linked to judgments about their teaching. “What if you come on a bad day?” So the practice of frequent short visits called for in our new evaluation system is crucial to countering the fear of infrequent high-stakes observations. But the quality of conversations that follow the frequent visits are the real metric teachers will use. Are they insightful? Are they helpful? This connects to another danger: inadequate attention to increasing evaluator expertise.

Evaluators will succeed in changing the culture around the term “needs improvement” when they are seen as constant learners about quality teaching themselves, become acute observers, are transparent about their own growth areas, and use data and questions as the basis of their conversations with teachers, not peremptory judgments. Further, when teachers have over time demonstrated progress in “needs improvement” areas, evaluators should recognize this success and the action steps that contributed to it.

We need more frequent and informal contact between evaluator and educator and more peer interaction. The culture of continuous improvement will be aided by a culture of regular communication and conversation about the work of educating. We need to highlight the right indicators, train evaluators to make better judgments, and create open-ended dialogue between educators.

### **3. Using the rubrics, procedures, and electronic devices to make superficial evaluation more efficient but not necessarily more substantive**

Rubrics can be used to make superficial evaluation more efficient. This is especially true if they are tied to hand-held data devices. Let’s repeat that. Rubrics can be used to make superficial evaluation easier and more efficient. We don’t want that.

That would happen if the rubrics were used as a portable check-list. Visit a class, check the boxes, and move on. That is not the Board’s or the Department’s intention, but it could happen if one isn’t careful. The DESE rubrics are not supposed to be used as an observation instrument. And there is no rational basis

to think they could be with any validity. Such a use would fly in the face of evidence-based observation and feedback to teachers.

The purpose of the state rubric with all the indicators and elements is to provide a uniform *foundation* for defining what good teaching and learning looks and sounds like. The DESE rubric is intended for use by educators for self-evaluation. It should also be used to inform evaluators about the range of things to look for. It should be used to ground development of reliability across evaluators. It can guide districts in the development of video and artifact exemplars. And it can be used as a potential map for choosing professional development topics. But it should not be used as an actual instrument for recording ratings of teachers either in observations or in summary evaluations.

A few nitty-gritty points to remember: for summary evaluation at the end of an evaluation cycle we are only required to fill out five lines: one for each standard and one as a final summary. Teachers do not have to be rated on each of the 16 “indicators” or on each of the 33 “elements,” only on the four standards and the overall summary rating.

#### **4. Succumbing to the Time Crunch**

Many evaluators in Massachusetts have too many teachers to evaluate already. Nothing in the new regulations and guidelines changes that. This is a structural problem: we do not have enough staffing or necessary budgets to provide the number of personnel necessary to ensure frequent evaluation of teachers. We do, however, have the ability to increase the frequency of 15 minute observations and quality feedback for all teachers if we 1) disconnect these visits from formal cycles of pre-conferences/post-conferences and formal documentation, 2) invest long-term in quality professional development for evaluators on how to observe for what is important, and give focused, useful feedback, and 3) supervise and support evaluators to be in classes daily. So let us use the new state guidelines and regulations as an opportunity to design new management systems that maximize feedback to teachers and minimize paperwork and formal procedures.

The DESE task force on evaluation also focused on better use of peers to increase useful feedback and productive conversations with teachers. Setting goals as a department or grade level is an example. The point is that we cannot operate exclusively on the old paradigm of every individual educator setting goals, and being evaluated with equal intensity every two years. There simply aren't enough evaluators with enough hours to go around. We suggest that evaluators prioritize their time and focus more on 1) non-PTS teachers, and 2) educators identified as needing improvement or unsatisfactory. This means we need to be creative in using peer evaluations to help proficient and exemplary educators continue to grow professionally and focus formal evaluations on the 20% of educators who need the most attention. Why not create incentives for the small

percentage of teachers rated exemplary to work with and evaluate the other proficient teachers (it is all about growth), leaving the evaluation of needs improvement, unsatisfactory and non-professional status teachers in the hands of principals and other administrators? That would be in keeping with the intent of the regulations.

## **5. Inadequate training and certification of evaluators for accuracy, significance, and reliability**

Many in our teacher corps doubt the competence of their evaluators to gather data on their performance. There are many outstanding observers amongst the administrators in the Commonwealth, but also many who are not. The skepticism of many teachers reflects the paucity of rigor and extended training of evaluators on their ability to evaluate and the previous lack of focus on evaluating the evaluators themselves! The required modules DESE has delivered this year are a good start, but quite inadequate to raise the level of expertise of our evaluators to where it needs to be.

Whereas the educator evaluation rubric is comprehensive, it is also abstract as all such teaching rubrics are. The “boxes” stop short of delivering concrete images of what the behaviors look and sound like; they do not provide samples of evidence that would back up a claim that the performance described in a “box” was present. For example, what does it look like and sound like when a teacher is effectively convincing students they “can master challenging material through effective effort, rather than having to depend on innate ability”? This work is left to the districts and is vital to good implementation.

Those images need to be developed in common across all educators in a district. This is more than statistical “reliability.” Evaluators can be reliably *wrong*, that is, they can be consistent with one another in making erroneous judgments about a rating. Accuracy means linking valid evidence with a rating; it precedes and grounds inter-rater reliability.

“Significance” is another crucial variable on which evaluators must be consistent. It means that the most important things are noticed. For example, if the student learning experience is not aligned with the objective, the students may be having fun but not learning much. That is more important than “student engagement” with the activity and is therefore a more significant item to pick up as an evaluator.

These qualities of accuracy, reliability, and significance will require extensive practice and training of evaluators far beyond the orientation activities currently offered. The skills required span literal note-taking, using multiple sources of evidence, evaluating the rigor of student work and teacher feedback.

In addition, educators want quality feedback and a lot more of it than they are getting. We hear that over and over again in interview studies of teachers, especially the younger generation of teachers (Coggshall et al., 2011). In order to have that frequent high-quality feedback happen, we need more evaluators visiting classes and we need focused professional development on an array of specific skills:

- a. Deep study of the knowledge base on teaching so we know what the boxes in the state rubric look and sound like at proficiency, both in terms of what teachers are doing and what students are doing
- b. Evidence-collecting skills and literal note-taking skills
- c. Coaching for evaluators on where to position themselves and what to write down during a class visit
- d. Skill for analyzing student tasks for quality and rigor
- e. Reporting skills that link Claims and Evidence with Impact on students
- f. Reliability across evaluators
- g. Differentiated Conferencing Skills, i.e., skills for fostering dialog and giving teachers useful feedback that builds an adult culture of constant learning and improvement in the school, not a fear of judgment and ranking

## **6. Failure to raise the bar for high-expertise teaching and our support for developing it.**

Raising the bar for teaching expertise may be the most difficult but also most significant challenge of the new evaluation system. It is a challenge that goes far beyond evaluation, because raising the bar also has implications for teacher prep, new teacher induction, professional development, and all the other levers of influence on teacher capacity.

Many skills called for in the new evaluation system have been absent from teacher prep and the professional development opportunities of most of our teacher workforce. Frequent, detailed non-judgmental feedback to students is one example. Insightful error analysis of student work is another. Convincing students to believe in malleable ability is yet another. Yet these are amongst the three highest leverage skills for improving student achievement (Hattie 2012).

The rubric in the new evaluation system defines a range of skills and a level of practice above the norm of what one sees in most classrooms. We've got a long way to go to meet the proficiency standards in that rubric, and meeting those standards is especially important for children of poverty who come to school behind academically.

This call to raise the bar has two ramifications: first, it calls for an emphasis on developing educator capacity rather than evaluative judgment. District leaders can use the evaluation rubric as a roadmap and tune it to focus on the elements

that will lead to the most significant improvements in teaching expertise for their district, not as blunt instrument to just highlight shortcomings.

Second, professional development and an adult culture of constant learning will be required over a number of years to raise the level of teaching in districts. While the resources for this PD aren't available in all districts at this time, this is a frustration we must deal with in a measured way. But the focus on adult professional culture is available to us now as the workplace engine for study and constant improvement.

It's a tricky message for administrators to internalize and operate from. They need to adopt an attitude that: "We're not where we should or could be, but it's no one's fault. Without blame, without recriminations, we will identify high-expertise teaching skills we haven't yet developed and put our shoulder to the wheel and plan fully to acquire them."

Raising the bar to the standards of this new rubric and its implied definition of proficient teaching challenges administrators to face low-performing teaching with more honesty, more courage and more skill. The recent reports (Anderson: NYT 3/1/13) of the consistent high percentage of teachers rated proficient or better across the US, unchanged by the last four years of attention to educator evaluation, says we *have not* raised the bar. Avoiding this situation in the Commonwealth will require an unprecedented rallying of courage and support by superintendents for their evaluators. That is worth a state-wide conversation and a high level conference.

### **III. IMPROVING TEACHING AND LEARNING IN THE CLASSROOM**

We'd like to leave readers with a final point: we've got to make educator evaluation part of something much bigger – the improvement of teaching across the board for everyone, even for the most experienced and competent among us. What is central to improving student achievement is the clear, shared image of what good teaching is. Our new regulations and rubric give us the impetus and the charge to do that with care and focus over the next few years. Once we have that clarity, we can use our common understanding of high-expertise teaching in *all* the arenas where leaders contact teachers.

## References

Weisberg, David, Susan Sexton, Jennifer Mulhern, David Keeling. *The Widget Effect*. New York: The New Teacher Project, 2009

Coggshall, Jane G., Ellen Behrstock-Sherratt, Karen Drill. *Workplaces That Support High Performing Teaching and Learning*. New York: American Federation of Teachers and American Institutes for Research, 2011.

Hattie, John. *Visible teaching for Teachers*. New York: Routledge, 2012

“Curious Grade for Teachers: Nearly All Pass” by Jenny Anderson in *The New York Times*, Mar. 31, 2013 (p. 1, 4), <http://nyti.ms/YUv9lt>