

Getting Teacher-Evaluation Rubrics Right

by Kim Marshall

Chapter 11 in *Rubric Nation: Critical Inquiries on the Impact of Rubrics in Education*, edited by Michelle Tenam-Zemach and Joseph Flynn, Jr. (Information Age Publishing, 2015)

Introduction

When I was first introduced to the idea of rubrics in a summer workshop with Grant Wiggins in the mid-1990s, the idea was immediately appealing: what a good way to get a handle on the perennially difficult task of evaluating students' written work. I was a principal of a large elementary school in Boston, and that fall we created grade-by-grade rubrics with three domains: Mechanics and Usage, Content and Organization, and Style and Voice. This was an extraordinarily helpful exercise, forcing us to make clear what writing looked like at four different levels of proficiency. For several years, we used the rubrics to evaluate students on a quarterly basis and inform and improve instruction.

I didn't become aware of teacher-evaluation rubrics until several years later. Charlotte Danielson published her *Framework for Teaching* in 1996, but a committee to revamp Boston's teacher-evaluation process, on which I served, did not adopt her work; we were still wedded to the narrative evaluation model with a two-point evaluation scale – Satisfactory and Unsatisfactory. We did create a list of desirable teacher attributes in each domain, which, in retrospect, was a baby step toward the development and use of rubrics for teacher evaluation.

Writing a teacher-evaluation rubric

In 2002, I left the Boston schools and started coaching principals. I also began writing *The Marshall Memo*, a newsletter summarizing educational research and best practices. Reading a wide range of publications, I saw more and more mentions of rubrics for teacher evaluation, and was drawn to the idea. I had been trained in the traditional approach to teacher evaluation and had written hundreds of lengthy narratives, and rubrics struck me as having the potential to be more time-efficient for supervisors and more informative for teachers.

In 2006, one of the principals I was coaching asked me to write a rubric for his charter school in Newark, New Jersey. I inquired why he did not just use Danielson's or one of the other rubrics that were available. He said he wanted to use something a little different and I accepted the job. Creating a teacher-evaluation rubric from scratch was an interesting challenge. I gathered several rubrics and saw a number of common flaws that I was determined to avoid:

- Most rubrics were long and wordy – very cumbersome for supervisors to fill out and teachers to digest in the busy world of schools.
- Some rubrics divided teaching into domains that were illogically conceptualized or too academic.
- Most used “fixed mindset” language for the levels (for example, *Distinguished*, *Proficient*, *Basic*, and *Unsatisfactory*), which could convey the idea that good teachers are born, not made.

- Some rubrics did not describe observable classroom behaviors in clear, vivid language.
- Some put two or more teaching behaviors into individual boxes, making it difficult for supervisors to give a single score for that portion of the rubric.
- Some underemphasized or completely left out important aspects of teaching (for example, homework, teacher attendance, and parent involvement).
- In most rubrics, each domain sprawled over two or more pages, making it difficult to see the overall picture of a teacher's performance in that area.
- Some rubrics used shortcuts that failed to capture important variations in teaching quality (for example, starting each of the four levels with words like *Always*, *Mostly*, *Sometimes*, *Never* and not providing detail on the gradations of classroom actions).

With these potential pitfalls in mind, I set about trying to create a better mousetrap.

The first decision any rubric-writer needs to make is how teachers' work should be divided up. After examining the available rubrics, research on effective teaching, especially *The Skillful Teacher* by Jon Saphier and Robert Gower (1997), *What Works in Schools* by Robert Marzano (ASCD, 2003), and *Linking Teacher Evaluation and Student Learning* by Pamela Tucker and James Stronge (ASCD, 2005), and insights from my years as a teacher, central office administrator, and principal, I decided on six domains:

- A. Planning and preparation for learning
- B. Classroom management
- C. Delivery of instruction
- D. Monitoring, assessment, and follow-up
- E. Family and community outreach
- F. Professional responsibilities

I thought these covered all of teachers' day-to-day responsibilities, were logically sequenced, and had about the right proportions – two-thirds focused on classroom instruction, one-third on aspects of the job outside the classroom.

The rubric writer's second decision is how many scoring levels there should be and how they should be labeled. Currently, virtually all teacher rubrics use a 4-3-2-1 scale, and I agreed with the logic of that approach. A four-point scale makes it possible to identify truly outstanding teaching (Level 4), solid professional performance (Level 3), mediocre practices (Level 2), and unacceptable teaching (Level 1). Tennessee educators are using a statewide 5-point scale, but interestingly, this rubric has written descriptions only at Levels 5, 3, and 1, leaving it to supervisors to "eyeball" Levels 2 and 4

(<http://team-tn.cloudapp.net/wp-content/uploads/2013/12/General-Educator-Rubric.pdf>).

The labels used for each level are also important; they embody the rubric's philosophy about performance and convey important messages to teachers and administrators. Over the last few years, there has been a shift from "fixed mindset" language for Levels 4 and 3 (*Distinguished*, *Exemplary*, *Excellent*, and *Proficient*) to "growth mindset" language (*Highly Effective* and *Effective*). Finding the right label for Level 2 is particularly tricky – how to describe mediocre performance without labeling and discouraging the teacher. Here are some

attempts: *Needs Improvement, Partially Effective, Minimally Effective, Basic, and Developing*. My thinking on labels has evolved, and I believe these are best:

4. *Highly effective*
3. *Effective*
2. *Improvement necessary*
1. *Does not meet standards*

Each is designed to convey general feedback about the teacher's performance, focus on standards and results, and encourage improvement.

Having made these decisions, I zeroed in on Level 3 (Effective), gathered descriptions of good teaching practices from multiple sources, sorted them into the six domains in my computer, arranged them in a logical sequence, eliminated duplication, and worked hard to make each one as clear, descriptive, and brief as possible (I kept each description to one line). Without question, this was the most intellectually demanding part of the entire process.

When my Level 3 descriptions were finished, I took each one up a notch to create Level 4, pegging them to the kind of outstanding teaching practices exemplified in Doug Lemov's (2010) book, *Teach Like a Champion*. I then took each Level 3 description down a notch to create Level 2 (decidedly mediocre practices that no teacher should be proud of), then took each one down another notch to create Level 1 (clearly unsatisfactory and ineffective practices). The final step was giving each row a one-word "headline" on the left side. [To see the current edition of the rubrics, got to www.marshallmemo.com and click Kim's Published Writing.]

The Newark charter school was pleased with the rubrics I created and proceeded to use them for teacher evaluation. With the principal's permission, I began to share the rubrics with other schools and wrote an article about them for Phi Delta Kappan *EDGE* magazine (2006). I decided to offer the rubrics as free and open source documents, and they are being used widely around the country and have been included on the New York and New Jersey lists of approved rubrics.

Over the last ten years, I have conducted hundreds of professional development workshops on how to make rubrics (not just mine) part of an effective teacher-evaluation process, and made thousands of classroom visits with school leaders with debriefings afterward. What follows are my insights on introducing rubrics to teachers and administrators and implementing them in ways that will maximize positive impact on teaching and learning. I will also address a number of commonly asked questions.

Introducing Rubrics

When teachers and administrators are first handed a copy of an evaluation rubric, the most common reaction is, "Holy cow!" (or words to that effect). Many educators are overwhelmed as they leaf through page upon page of densely packed matrices and hardly know where to begin. I have found that the most effective way for a superintendent, principal, or consultant to get teachers to move past this initial negative reaction and see the benefits of a rubric is to proceed in a step-by-step fashion, alternating between a broad overview (general) and granular specifics (particular). Here is a sample rubric introduction using this approach:

- *Step 1 (General)* – Briefly review the rubric’s domains and rating scale. This gives people a sense of how the document is structured.

- *Step 2 (Particular)* – Have everyone turn to one page (I suggest domain D, Monitoring, Assessment, and Follow-up) and silently read the vertical column at Level 3. When people have finished, pick one cell at Level 3 and read it aloud (for example, “Has students set goals, self-assess, and know where they stand academically at all times”). This demonstrates that Level 3 is a solid, expected professional performance – nothing that any teacher should be ashamed of – which pre-empts one of the most common pushbacks: teachers’ resistance at getting a rating that feels like a B.

- *Step 3 (Particular)* – Read aloud the horizontal row from which the Level 3 example was taken, moving from right to left. For example:

Level 1: Allows students to move on without assessing and improving problems with their work.

Level 2: Urges students to look over their work, see where they had trouble, and aim to improve those areas.

Level 3: Has students set goals, self-assess, and know where they stand academically at all times.

Level 4: Has students set ambitious goals, continuously self-assess, and take responsibility for improving performance.

This gives people a sense of how the language of the rubric moves from unsatisfactory to excellent.

- *Step 4 (General)* – Display that page of the rubric on a screen and read the headings of each horizontal row to give an overview of the areas covered in that domain – for example, Criteria, Diagnosis, On-the-spot, Self-assessment, Recognition, Interims, Tenacity, Support, Analysis, and Reflection.

- *Step 5 (Particular)* – Display that page of the rubric with the ratings of a sample teacher circled or highlighted. A winning strategy is to show how you would rate *yourself* as a teacher (the more mixed your self-assessment, the better). This shows the audience what a realistic evaluation looks like – and models humility.

- *Step 6 (General)* – Display the final summary page of the rubric, again with a sample teacher’s ratings in all the domains circled or highlighted. Again, it’s helpful if this is your own self-assessment.

- *Step 7 (Particular)* – Now have people think of a teacher they know well and silently rate the teacher on all the lines on that page. Emphasize that it’s important to read Level 3 of each row first, and if that doesn’t describe the teacher’s performance, look left or right for the best description. A room full of educators usually gets very quiet as people fill out the page.

- *Step 8 (General)* – When everyone is finished (this usually take about 3-4 minutes), ask if any teacher got all Level 4 ratings. It’s very rare for there to be a single hand in the air. Ask if there was at least *one* Level 4 rating. Several hands usually go up, which makes the point that in any school, there will be a few top ratings, but not a lot.

After Step 8 is a good time to invite people to discuss in small groups the possible advantages and disadvantages of using rubrics to evaluate teachers. After five or so minutes of group discussion, reconvene the audience and have participants share their thoughts. Begin with advantages – this is an important tactical move to prevent skeptics from setting a negative tone up front. These are some of the positive points that are usually mentioned by workshop participants:

- Rubrics provide a shared conceptual framework and a common language about good and not-so-good teaching.
- Rubrics tell teachers exactly where their performance in each area stands on a 4-3-2-1 scale.
- Since most teachers naturally aspire to the highest level, a four-level rubric has a built-in push toward excellence.
- For teachers with scores at Levels 3, 2, and 1, the language of the cell just above where they scored provides a specific description of what they need to do to improve.
- The comprehensive scope of a rubric makes it easier for a teacher to accept criticism (Level 2 or 1 ratings), since there are many other areas in which they are (hopefully) scoring at Level 3 and 4.
- The rubrics can easily be used by teachers to self-assess and set goals for a school year, and then track their progress over time.
- Rubrics are much quicker for supervisors to fill out than traditional narratives, since the writing is done for them.

Turning to the disadvantages of rubrics (and it is important to air critical feedback and apprehensions), these are some of the concerns that are often voiced:

- If supervisors have not visited classrooms frequently, they will not be able to fill out the rubrics knowledgeably.
- If supervisors have observed only “glamorized” lessons put on for their benefit, their rubric scoring will not reflect the kind of teaching students are experiencing on a daily basis, which is what really matters for long-term learning.
- The rubric might limit supervisors’ perceptions and prevent them from seeing and commenting on important aspects of instruction.
- Some supervisors might not be considered credible evaluators in certain subject areas or grade levels.
- Some administrators’ might be uncomfortable judging teachers on a 4-3-2-1 scale, or uncertain of their ability to make fair judgments in so many areas.
- Supervisors might not be fully candid and refrain from giving Level 2 or 1 scores to teachers who deserve them, undermining the potential for useful feedback.
- Conversely, some administrators might decide to give very few Level 4 scores, “grading on a curve” irrespective of the amount of excellent classroom performance in their school.
- Supervisors might complete the rubrics and present them to teachers without providing the teacher the opportunity to voice disagreements.

- Some educators do not like this type of precise, pre-packaged evaluation tool. As one private-school administrator in Washington D.C. said to me, “I don’t like being put in a box.”
- Teachers might count up and compare points with colleagues (4 for each top-level rating, 3 for each next-to-top, etc.), turning the evaluation process into a numbers-driven competition to gain an advantage over their colleagues.
- Supervisors might require teachers to provide evidence for each rating, which makes filling out the rubric extremely time-consuming and can distract teachers from their work with students (a Connecticut teacher who was asked to do this in late 2013 said, “I’ve never worked so hard and taught so little.”)
- Making rubric scores high-stakes – for example, giving merit pay bonuses for Level 4 performance or moving to dismiss teachers scoring at Level 1 – might make it less likely that teachers would be candid about their shortcomings and work effectively with their supervisor.

In my experience, this kind of pro-and-con discussion almost always leaves an audience of teachers and school administrators with a positive feeling about using rubrics – along with some legitimate concerns. And indeed, the devil is in the details. Even with all their potential benefits, rubrics can be implemented in ways that damage staff morale and do little or nothing to improve teaching and learning. As I have learned over the years, thoughtful implementation is as important as the quality of the rubric itself.

Hurdles to Successful Implementation

The most important question is when and how teacher-evaluation rubrics should be used. Back in 2006, my assumption was that rubrics were summary evaluation tools and supervisors would wait until the end of the school year to fill them out. But teachers naturally want to know where they stand month by month, and many superintendents want to keep track of how principals are handling the evaluation process as the year unfolds. This hunger for real-time evaluation data has led many supervisors to fill out rubrics throughout the year – either during classroom visits or in post-observation conferences with teachers.

Using rubrics in this manner appeals to busy school leaders since it makes the teacher-evaluation process seem quicker and more efficient. However, I believe real-time rubric scoring undermines effective coaching, supervision, and evaluation. For starters, it is next to impossible to fill out a comprehensive, multi-page rubric while observing the fluid and complex dynamics of a classroom. To be good observers, supervisors need to have their heads up, walk around the room, listen carefully to the teacher, look over students’ shoulders at the work they’re being asked to do, ask one or two students “What are you working on?,” and look at some of the material on the walls. The challenge – and it is challenging – is to decide on the most important thing to address with the teacher afterward and jot down a few quick notes. What affirmation, question, or concern is most likely to move the teacher’s practice forward? Trying to rubric-score a teacher during an observation distracts the supervisor from focusing on those key insights and makes his or her feedback to the teacher seem bureaucratic and inauthentic.

Regardless, a number of software companies (among them Rally, Teachscape, and iObservation) have created electronic applications that allow supervisors to fill out rubrics on iPads or laptops during observations and even send them electronically to teachers before leaving the classroom. While these products are highly seductive to school leaders, I believe there are several practical problems: (a) the supervisor is making snap judgments about what is happening in the classroom without giving the teacher a chance to explain the bigger picture; (b) the teacher's anxiety level is likely to spike knowing that the supervisor is making evaluative comments in electronic form in real time; and (c) it's less likely that there will be a follow-up conversation because it appears that the supervisor has already made up his or her mind. This is a perfect example of one of Charlotte Danielson's (2007) most telling critiques of conventional evaluation: It is done *to* teachers, not with them (, p. 182).

Some proponents of in-class rubric scoring (whether it be through using an app or traditional paper and pencil) acknowledge the problem of supervisors being overwhelmed by how much happens during a classroom visit and suggest that only one segment of the rubric be filled out (perhaps the school has decided to focus on teacher's questioning strategies or classroom management). But this seriously limits the supervisor and might lead him or her to miss important interpersonal or pedagogical events. And even with only one page of the rubric to evaluate, there is still the problem of finding the right line in the rubric and making immediate evaluative judgments while so much is occurring simultaneously.

Filling out the rubric during post-observation conferences with teachers creates a different set of problems. First, rubrics are inherently evaluative, and most principals I am working with find that scoring teachers on a 4-3-2-1 scale after each observation undermines productive coaching since the teacher tends to remember only the ratings. Second, receiving evaluative feedback on twenty or thirty areas is overwhelming, and the result is likely to be a stressed-out teacher with no clear focus for improvement. Third, being required to use rubric language after a classroom visit may prevent supervisors from articulating in their own words the one or two most important commendations or suggestions that the teacher needs to hear.

What Is to Be Done?

I have argued that using rubrics during classroom visits or immediately afterward is problematic. But waiting until the end of the year to fill out the rubric runs the risk of teachers not receiving timely feedback on their performance and possibly getting blind-sided by negative feedback in the final evaluation meeting. The solution, I have come to believe, is using rubrics formally at three points in the year:

- As school opens, teachers score themselves on the entire rubric, meet with their supervisor, and set 2-3 improvement goals based on areas of the rubric with relatively low scores;
- At mid-year, teacher and supervisor meet and compare the teacher's updated self-assessment with the supervisor's tentative scoring page by page, discuss any differences, assess progress on the teacher's goals, and identify areas for growth and support;

- At the end of the year, teacher and supervisor repeat this process and reach closure on the year's ratings.

I have found that the mid-year and year-end evaluation meetings work best if the teacher and supervisor fill out the rubrics beforehand (in pencil), discuss only the areas where they disagree (and perhaps one or two other areas that are particularly important to the teacher or supervisor), and base the conversation on the teacher's actual performance (versus second-hand information or philosophical viewpoints). This approach makes rubric meetings quick, focused, and efficient – usually not more than 30 minutes.

As the year-end meeting proceeds, the supervisor provides an overall rating for each domain (with a brief comment if needed). When all the domains have been scored, the supervisor fills out the final summary page, gives an overall rating, they both write brief summative comments, and they sign off.

The only exception to this three-times-a-year process is with teachers whose performance shows clear signs of being unsatisfactory (Level 1). As soon as serious performance problems become apparent, these teachers should receive scores on the relevant areas of the rubric, an improvement plan, and intensive support.

Frequently Asked Questions

When schools and districts implement teacher-evaluation rubrics, a number of questions come up. Here are my thoughts on a selection of these:

- *How many classroom visits are needed to fill out the rubric?* I have long believed that the traditional model (announced full-lesson observation with pre-conference, lengthy write-up, and post-conference) has built-in design flaws that make it inaccurate (the supervisor is not getting a sense of what students are experiencing on a day-to-day basis), ineffective (not seeing typical daily practice, the supervisor can't give appropriate coaching suggestions), and dishonest (if lay people knew that teachers were being evaluated on an annual "dog-and-pony show", they would be scandalized). In addition, the traditional process consumes hundreds of hours of supervisors' time. Instead, I believe that short, frequent, unannounced classroom visits followed by brief face-to-face post-observation conversations (with very brief write-ups afterward) allow supervisors to sample day-to-day instruction, have regular coaching conversations, and gradually gather most of the information they need to complete the rubric at year's end (Marshall, 2006).

How many short observations there should be is open to debate; my experience, as a Boston principal and as a coach of principals, is that ten visits per teacher per year is about the right frequency (Marshall, 2013). In most schools, this means two or three classroom visits a day. Some districts are asking supervisors to make fewer visits (sometimes combined with one or two full-lesson observations). Other educators advocate more-frequent visits – for example, Newark, New Jersey charter school leader Paul Bambrick-Santoyo (2012), in his book, *Leverage Leadership*, makes the case for *weekly* classroom visits for each teacher.

It is difficult for school leaders to squeeze more than two or three observations into their incredibly busy days, and we need empirical research on how many visits provide a reasonably accurate picture of a teacher's performance – and understand whether there is a point of

diminishing returns on classroom observations. Equally important, in my experience, is supervisors having a good “eye” for instruction, genuinely listening to teachers’ insights when they talk to them after each visit, and, when necessary, reaching out to content-area instructional coaches and subject-area specialists for guidance.

- *How can teachers’ professional work outside the classroom be evaluated?* This includes extra duties, paperwork, attendance and punctuality, working with colleagues on curriculum planning and data analysis, professional development workshops and courses, book study groups, etc. To assess these important areas, supervisors must rely on other points of contact with teachers: dropping in on team meetings, observing staff meeting interactions, attending student performances and athletic events, watching teacher interactions with parents, and monitoring data on attendance, paperwork, and other areas. Supervisors who are visible around their schools can form a fairly accurate picture of teachers’ non-classroom performance, but there will inevitably be gaps in their knowledge. This is where teachers’ input in mid-year and end-of-year rubric conversations is so helpful. It’s also a good idea to have more than one administrator monitoring teachers’ performance so the leadership team can compare notes and fill in gaps.

- *In evaluation conferences, who has the burden of proof?* When a supervisor and teacher disagree on a particular line of the rubric, here is a possible rule of thumb: if the teacher is advocating for a Level 4 score and the supervisor believes Level 3 or 2 is more accurate, the teacher needs to convince the supervisor. If the supervisor is arguing for a Level 1 score but the teacher believes a higher score is more accurate, the supervisor needs to have evidence. If there are disagreements between Level 2 and 3 scores, there should be a free-flowing debate on the evidence. At the end of the day, however, the supervisor has the final say.

- *How much are teachers and supervisors likely to differ when they compare ratings?* In one New York City elementary school that used this approach during the 2012-2013 school year, the principal reported to me that 60% of teachers’ self-assessments were virtually identical to her ratings, 30% percent of teachers rated themselves lower, and 10% scored themselves significantly higher than the principal believed they deserved (personal communication, June 2013). She was glad she had done mid-year check-ins, because it offered her time to address these discrepancies in her classroom visits and coaching conversations.

- *Should novice teachers be evaluated with a modified rubric?* The rationale for creating a less-demanding rubric is that new teachers will not receive many ratings at Level 3 and 4 on the standard rubric and might become discouraged. However, new teachers are working with students every day, and I believe they should be held to the same expectations as everyone else. Administrators should explain that it is not the end of the world for new teachers to have some Level 2 ratings (even a few at Level 1) in the first year; the school will provide lots of support to put them on a steep learning curve so they reach the effective level in virtually all areas in their second and third year of teaching.

- *Should fractional scores be permitted?* For example, if a teacher’s performance in one area straddles Level 2 and 3, is it appropriate to give a 2.5? Since secondary teachers work with four or five different groups of students and elementary teachers cover several different subjects,

this seems like a logical approach. On the other hand, there is an argument for keeping things simple and asking supervisors to use only whole numbers based on the preponderance of the evidence. A compromise might be to allow supervisors to straddle ratings in the mid-year conference (which would give the teacher a clear signal to step up performance in a particular area) but require whole-number ratings in final evaluations.

- *How should teachers' absences be counted?* Not all rubrics have a line for attendance, but for those that do (in my rubric, Level 4 is 98-100% present, Level 3 is 95-97% present, Level 2 is 6-10% absence, and Level 1 is 11% or higher absence), should teachers be considered absent for taking personal days, attending a professional conference or visiting another school? My recommendation is that only sick days be counted as absences, and there should be a place to note if there are exceptional circumstances that explain unusually low attendance – for example, a serious illness or a death in the immediate family. Some teachers disagree with a percentage approach to attendance, but I think it is important that the rubric sends a clear signal that a teacher's presence in the classroom is critical – substitutes almost never deliver instruction that is as rigorous and effective as the regular teacher (Clotfelter, Ladd, & Vigdor, 2007; Miller, Murnane, & Willet, 2007).

- *How should non-classroom school staff be evaluated?* These include librarians, nurses, guidance counselors, instructional coaches, and other non-teaching roles. Charlotte Danielson (2007) has published a series of rubrics for non-teaching staff, and the Westwood, Massachusetts schools have done similar work branching off from my teacher rubrics. For educators who work with students in confidential settings (e.g., psychologists and guidance counselors), direct observation is not possible and supervisors need to schedule regular check-in meetings and perhaps use anonymous student surveys to assess the quality of the work.

- *Should teachers' ratings be made public?* In my view, absolutely not. I believe rubric scores are confidential personnel records and should not be released, even in aggregated form at the building level. Public disclosure of rubric scores not only violates the trusting relationship that must be nurtured between supervisors and teachers, but also runs the risk of putting pressure on supervisors to inflate scores (who wants their school to look bad in the community?). However, principals and superintendents should create a confidential spreadsheet of teachers' ratings for their leadership teams, highlighting the areas where there is solid performance and where additional work and professional development are needed.

- *Should top-rated teachers receive merit pay?* My reading of the research on performance pay is that even in the business world, it's not an effective practice, and a few experiments in K-12 education have not produced positive results and have been discontinued, sometimes for financial reasons (Johnson & Papay, 2010; Springer & Gardner, 2010; Frey & Osterloh, 2012; Yuan, K. et al, 2013). However, there are several ways that rubric evaluations might be used to tweak the traditional salary scale:

- Top-rated teachers might be offered opportunities to perform extra work for extra pay – for example, mentoring colleagues, designing curriculum, and running professional development.

- Districts might consider eliminating or scaling back salary increments for advanced degrees, since the research shows very little correlation between master's and doctoral degrees and classroom effectiveness (Hattie, 2007; Goldhaber, 2002). For example, Hattie's (2007) meta-analyses rank teacher credentials 124th out of 138 in overall impact on student achievement. The money saved could go to compensating highly effective teachers for additional duties.
- Teachers scoring at Level 2 might be denied a step increase the following year and receive an improvement plan and intensive support; if they have not progressed to Level 3 by the end of that year, they might be dismissed. The Hillsborough, Florida school district used this approach with union approval; the union president said she would not want her own children in a mediocre classroom and didn't think anyone else would either (Wingert, 2010).
- Teachers evaluated at Level 1 should get an immediate improvement plan and intensive support and, if they do not improve in a reasonable period of time (preferably within that year), should be counseled out or dismissed.

This approach allows rubric ratings to play a part in teachers' income without using the ineffective strategy of merit pay.

• *Should teachers' rubric scores be recorded in numerical form?* One built-in feature of rubrics is that it's easy to tabulate teachers' performance in a very precise manner; for example, Valerie Williams scored 29 out of a possible 40 in classroom management. But there are several problems with this approach. First, adding up ratings assumes that each line in the rubric has equal value, which is not true. The research is clear that some teacher actions have more impact on student learning than others. For example, Hattie's (2008) comprehensive compilation of meta-analysis data ranks effective use of formative assessment data third (out of 138) in its impact on student achievement, while homework is ranked 88th. Using numerical rubric scores conveys a false sense of precision to teachers and school leaders.

Second, certain parts of the rubric are more important for some teachers than others. Perhaps Valerie set a personal goal of working with her grade-level team on interim assessment follow-up, developing a class website, and communicating with hard-to-reach parents by text messages; for her, those rubric areas have more weight than areas in which she is doing fine. Third, numerical scores do not capture a teacher's growth in specific areas. Perhaps Valerie and her team have made great strides in the quality of their unit plans and use of Essential Questions, meriting a special commendation from the principal, but only a one-point rise in her rubric score.

Fourth, a teacher might perform quite differently depending on the challenge level of students and the working conditions within a school – for example, teaching an AP history class versus a group of ninth-grade repeaters. The context in which a teacher teaches has tremendous impact on rubric scores, and supervisors need to take that into account, not by grading on a curve, but by making sure that nobody over-emphasizes the importance of numerical scores.

Finally, reducing rubric ratings to numbers might very well lead teachers to compare their evaluations and cause morale problems, especially since, as was argued above, the numerical scores do not reflect precisely how a teacher is performing in a particular year.

For these reasons, I think summative rubric evaluations should be reported using only the labels for each domain. For example, a given teacher's work was *Effective* in Planning and Preparation, *Highly Effective* level in Classroom Management, and so on, with an overall rating (e.g., *Effective*) summing up all the domains, accompanied by brief written comments where appropriate.

• *Will rubrics improve teaching and learning?* This, of course, is the most important question of all. It could be argued that rubrics are old wine in new bottles – merely a streamlined way of delivering top-down evaluative judgments that will not improve much of anything. That's possible, but I believe well-constructed rubrics, if used appropriately by competent supervisors, can add significant value. The key factors are a skillful introduction of the rubrics to teachers, teachers self-assessing and setting goals, well-trained supervisors making frequent classroom visits with feedback conversations and coaching throughout the year, and teacher input in mid-year and year-end evaluation meetings. In schools with these factors in place, rubrics bring much greater clarity to teacher supervision and evaluation, push educators to higher levels of performance, and provide a powerful boost to students' achievement and life chances.

Developing a Research Agenda for Rubrics

Since rubrics are a relatively new development in K-12 schools, there are a number of unanswered questions. that would be helpful for researchers to address in the years ahead. While this is not an exhaustive list, the following represents unanswered questions about teacher evaluation rubrics:

- Under what circumstances will the use of rubrics improve teacher evaluation and teacher learning?
- What is the ideal length and level of detail for a teacher-evaluation rubric (i.e., number of domains, number of lines per domain, total number of pages)? Similarly, what is the ideal number of rating levels?
- What kinds of professional development best prepares principals and teacher leaders to use rubrics effectively?
- What is the optimal supervisor/teacher ratio for effective supervision and evaluation?
- How and to what extent do electronic devices (e.g., iPads, laptops, smartphones) enhance the supervision and evaluation of teachers and when do they impede it?
- When during the school year do teacher evaluation rubrics best promote improvements in teaching and learning?
- Should teachers self-assess on rubrics? If so, at what point or points in the school year? And what is the most effective role of supervisors in this process?
- What is the value of releasing teachers' rubric scores to the public and does the practice have an impact on teaching and learning?

Teaching is an exceptionally complex enterprise, and evaluating teachers fairly and accurately has always been a challenge. Considering that the average instructor teaches five lessons a day, which adds up to 900 lessons a year, and the teachers in a medium-size school collectively teach about 27,000 lessons a year, the job of supervision and evaluation seems

impossible – especially with the traditional evaluation process. Teacher-evaluation rubrics have great potential in terms of building a shared understanding of good teaching and streamlining the process – *if* they are well-constructed, *if* they are introduced to teachers in a way that builds ownership and trust, and if they are used wisely by thoughtful and well-trained administrators who regularly visit classrooms and follow up with teachers.

References

- Paul Bambrick-Santoyo (2013). *Leverage Leadership* (San Francisco, CA: Jossey-Bass.)
- Charles Clotfelter, Helen Ladd, & Jacob Vigdor (2007) “Are Teacher Absences Worth Worrying About In The U.S.?” NBER Working Paper No. W13648, <http://www.nber.org/papers/w13648>
- Bruno Frey & Margin Osterloh, “Stop Tying Pay to Performance” in *Harvard Business Review*, January/February 2012 (Vol. 90, #1-2, p. 51-52)
- Goldhaber, D. (2002). The mystery of good teaching. *Education Next*, Spring, 2002, p. 50-55. Retrieved from <http://educationnext.org/the-mystery-of-good-teaching/>
- John Hattie (2008). *Visible learning. A synthesis of over 800 meta-analyses relating to student achievement*. New York, NY: Taylor and Francis.
- Susan Moore Johnson & John Papay, “Merit Pay for a New Generation” *Educational Leadership*, May 2010 (Vol. 67, #8, p. 48-52)
- Doug Lemov (2010). *Teach Like a Champion*. San Francisco, CA: Jossey-Bass.
- Kim Marshall (2006). “The Why’s and How’s of Teacher Evaluation Rubrics” *EDge Magazine*, September/October 2006, Vol. 2, #1, p. 1-19)
- Kim Marshall (2013). *Rethinking Teacher Supervision and Evaluation* (San Francisco, CA: Jossey-Bass, second ed.).
- Robert Marzano (2003). *What Works in Schools*. Alexandria, VA: ASCD
- Raegen Miller, Richard Murnane & John Willett (2007). “Do Teacher Absences Impact Student Achievement? Longitudinal Evidence from One Urban School District”, NBER Working Paper

No. W13356, <http://www.nber.org/papers/w13356>

Jon Saphier & Robert Gower (1997). *The Skillful Teacher*. Acton, MA: Research for Better Teaching.

Matthew Springer & Catherine Gardner, "Teacher Pay for Performance: Context, Status, and Direction (2010). *Phi Delta Kappan*, May 2010, Vol. 91, #8, p. 8-15.

Pamela Tucker & James Stronge (2005), *Linking Teacher Evaluation and Student Learning*. Alexandria, VA: ASCD

Pat Wingert, "When School Reformers and Union Leaders Unite", *Newsweek*, December 6, 2010, <http://www.newsweek.com/when-school-reformers-and-union-leaders-unite-68989>

Kun Yuan, Vi-Nhuan Le, Daniel McCaffrey, Laura Hamilton, Brian Stecher, Julie Marsh, and Matthew Springer, "Incentive Pay Programs Do Not Affect Teacher Motivation or Reported Practices: Results from Three Randomized Studies", *Educational Evaluation and Policy Analysis*, March 2013 (Vol. 35, #1, p. 3-22)

This article gives details on the Florida legal challenge:

Kathleen McGrory and Danny Valentine, "Teachers Mount Legal Challenge to New Evaluation System" in *Tampa Bay Times*, April 16, 2013.