

Multiple Measures of Teacher Effectiveness in Hillsborough County Public Schools

Hillsborough County Public Schools (HCPS), the eighth-largest school district in the United States with 191,000 students, had a problem. Like most public school districts in the nation, too many of its teachers were rated outstanding. In 2009, principals gave 46 percent of high school teachers perfect scores on the district's evaluation rubric. The pattern followed a nationwide trend; a recent study of 12 districts in four states found that less than one percent of teachers received a negative evaluation rating.¹ In districts using more than two rating levels, 70 percent of tenured teachers received a perfect evaluation score.² Yet, student achievement on state comprehensive assessments did not reflect the outstanding reviews.³ In HCPS, the district's diverse student population – 42 percent white, 22 percent black, 28 percent Hispanic, and 57 percent qualifying for free or reduced priced lunch – showed moderate improvement in reaching math and reading grade-level proficiency. But, progress had stalled, and the proficiency gap between white and Asian students and their black and Hispanic peers persisted (see **Figures 1 and 2**).⁴ Vague evaluation systems applied inconsistently across school districts like HCPS produced invalid and unreliable assessments of teachers' performance. Specifically, the overwhelmingly positive evaluation results did not correspond to rising student achievement. More problematic, the evaluations offered no actionable feedback, limiting teachers' professional growth.

Superintendent MaryEllen Elia identified these and other problems with the district's teacher evaluation system early in her tenure as leader of the Tampa, FL-based district. By 2008, Elia and her leadership team were working closely with the teachers' union, the Hillsborough Classroom Teachers Association (HCTA), and its president Jean Clements, on reforming the way teachers were evaluated. They believed any new system should have multiple measures of teacher effectiveness that balanced principal and peer observations with student achievement data. In 2009, HCPS established a Teacher Evaluation Committee comprised of 28 members, including principals and teacher representatives from the HCTA, to design a new evaluation system. Soon after, the district was selected as an Intensive Partnership for Effective Teaching site by the Bill & Melinda Gates Foundation. The collaboration allowed Elia, Clements, and their staffs to design and implement a comprehensive evaluation system much more quickly than previously imagined.

By the spring of 2010, the Teacher Evaluation Committee had drafted a new teacher evaluation system premised on multiple measures of teacher effectiveness. Peer and mentor observations, principal observations, and student achievement data formed the system's backbone. Observations from a peer (for experienced teachers) or mentor (for new teachers) evaluator accounted for 30 percent of a teacher's evaluation; another 30 percent came from the principal. The remaining 40 percent integrated student achievement data using value-added measures calculated by the University of Wisconsin's Value-Added Research Center. Performance management software, pre- and post-observation conferences, and informal classroom "pop-ins" provided additional feedback and data sources. With the new teacher evaluation, the role of the principal also dramatically changed, leading HCPS to develop a new principal evaluation system. Principal evaluations were revised to incorporate teacher feedback and consistency between teacher evaluation scores and student achievement data.

Extensive training began in the summer of 2010 to prepare principals, assistant principals, and newly hired peer and mentor evaluators for the system's rollout in the 2010-2011 school year. HCTA was an involved and supportive partner throughout the process, part of what Teacher Evaluation Committee Co-Chair Becky Kaskeski called an "extraordinary relationship" that had been and remained "very collaborative." And as evidence of the deep district-union collaboration, 96 percent of HCTA's members who voted approved a new contract including the multiple-measures evaluation model. HCPS's efforts were only one approach to developing fairer and more reliable measures of teacher performance. Yet, with peer, principal, and value-added components developed in partnership with the teachers' union, the multiple-measures evaluation had the potential to serve as a model for the national effort to improve the way teachers were evaluated.

A New Evaluation Rubric

At the foundation of the peer and principal evaluations was a new observation rubric based on teaching authority Charlotte Danielson's work. The old evaluation system had three performance ratings with an opportunity to earn "outstanding" marks in a few instructional-focused categories. But, the distinctions were rarely used; nearly all teachers met standards and many received outstanding ratings. As Tracye Brown, a member of the Teacher Evaluation Committee, explained, "It would appear that all of our teachers were high performing because the bulk of them were being rated outstanding, so there was very little differentiation in the scoring evaluation." The old evaluation model also did not provide useful feedback for teachers who wanted or needed to improve. According to Brown, this was one reason teachers requested a new evaluation system: "Our former evaluation system was not very descriptive and specific and we knew that we had the desire to move in the direction of a rubric that would help teachers identify effective behaviors as well as help principals in evaluating teacher behaviors."

The committee was drawn to the observation rubric created by Danielson, as it provided a tested, non-ideologically based and broadly applicable set of criteria for observations. Danielson met with the committee several times throughout 2009 to help craft a rubric and evaluation system that worked for the teachers and the specific needs of HCPS. The Teacher Evaluation Committee ultimately adopted Danielson's four-domain four performance-ratings rubric with a few slight adaptations (see **Figure 3** for an excerpt). Wording on the performance ratings was changed from "Unsatisfactory," "Basic," "Proficient," and "Distinguished" to "Requires Action," "Developing," "Accomplished," and "Exemplary," respectfully. In particular, teachers on the Evaluation Committee felt that the lowest rating of "Unsatisfactory" was too judgmental, while "Requires Action" accurately described what needed to happen if a teacher received the rating in a component. According to Kaskeski, "After much discussion, it was changed to Requires Action. Requires Action just means, 'I need some help on that.'"

The Teacher Evaluation Committee also added weights to Danielson's four domains: "Planning and Preparation" (weighted 20 percent), "The Classroom Environment" (20 percent), "Instruction" (40 percent) and "Professional Responsibilities" (20 percent). Each domain covered five to six components, for a total of 22. While 22 was a significant reduction from the previous evaluation's 41 components, HCPS leaders felt that the smaller number captured more and better data. With 41 components, the former teacher evaluation became a matter of checking boxes, rather than providing substantive and actionable feedback. The new components allowed peers and principals to gather actionable feedback on teachers' behaviors. For example, the Instruction domain measured a teacher's communication with students, use of questioning and discussion techniques, engagement of students in learning, the use of assessment in instruction, and the demonstration of flexibility and responsiveness. The first three domains were filled out by the peer or mentor teacher, and the fourth

domain was the exclusive responsibility of the principal. Members of the Teacher Evaluation Committee felt that only a principal could evaluate a teacher's performance in Professional Responsibilities, which included tasks such as maintaining accurate records and communicating with families. According to Kaskeski, "Only the principal is evaluating domain four. Do they come to work on time? Do they participate in special learning communities? Do they do all those things? The peer would have no way of knowing." Finally, Danielson reminded members of the Teacher Evaluation Committee that few teachers consistently scored Exemplary in a domain. According to Tricia McManus, co-chair of the Teacher Evaluation Committee, "The shift has been to understand that we are not all exemplary and that's one thing that Charlotte says about her rubric, exemplary is not a place that teachers live. They visit it depending on the lesson, depending on the day, but that every single lesson every single time is not going to be rated exemplary."

The new ratings and domains addressed a number of weaknesses in the old evaluation rubric. First, they accounted for the full dynamic of learning in the classroom. Kaskeski described the emphasis on learning articulated by the new rubric: "It has changed the focus from sole teacher behavior to the interaction that goes on. Do the kids know what they're supposed to be learning? Can they walk out of (the classroom) and tell you this is what I learned and this is how it applied to yesterday and this is what we're going to be doing with it?" The rubric also forced peers, mentors, and principals to make a definitive assessment about a teacher's performance in a specific area. There was no middle ground. According to Kaskeski, "We knew we wanted to have four levels; that forces people to pick something. You know if you have a five—five category or a three category — people seem to go right down the middle."

Peer and Mentor Evaluators

The peer and mentor evaluators were a crucial part of the new evaluation process, as subject-specific and experienced teachers assessed their peers' performance and aided development. Brown described how the idea of peer and mentor evaluators gained traction: "What we found from focus groups and conversations with teachers was a strong desire to have an expert, someone in their content and level, evaluate them—a peer." The idea of peer support was not entirely new to HCPS. According to HCTA President Clements, a peer assistance program with no evaluative component had been developed and approved by the union and school board in the 1990s. However, it never got off the ground due to budget constraints. Now, with support from the Gates Foundation, HCPS had the resources to fund the start-up costs of a peer observation and evaluation system.

Of course, observations and evaluations were only as good as the person conducting them. The deployment of respected and trained peer and mentor evaluators was critical to the success of the program. In the spring of 2010, more than 650 teachers applied for 117 peer and mentor positions. Hopefuls for a position completed a paper screening, including essay questions, and a screening committee comprised of principals, teachers, members of the HCTA, and district administrators screened the applicants, selecting individuals for interviews. That 11-member committee then conducted a 30-minute interview with applicants who passed the initial screening. Importantly, the interview included a role-play where prospective peers and mentors watched a video of a classroom teacher, taking notes on the teacher's strengths and weaknesses, and discussing with the committee — still in the role-play mode—the feedback they would deliver to that teacher. An important characteristic of a peer evaluator, according to Stephanie Woodford, who oversees the peer evaluators for the district, was the ability to take ethical, honest "snapshots" of a teacher, as well as the ability to connect with teachers, build rapport, and demonstrate leadership.

In addition, successful peer and mentor teachers were willing and able to learn the spectrum of content areas his or her observations covered. Woodford explained: “An elementary peer has to become an expert in every single area of content for elementary school because you’re going into a second grade classroom, but it might be a writing lesson, reading, or math. So you have to know all of that content.” That subject familiarity allowed peers to make insightful evaluations on the lessons observed, and it also facilitated specific and meaningful feedback to the teacher. So far, Woodford has noticed an encouraging rapport between teachers and peer evaluators, one that continues past the observation cycle: “What we’re finding is it carries on after the observation. The teachers continue to e-mail the peers if they had a positive experience, asking them for more suggestions and things.”

Over the summer of 2010, the peers, mentors, assistant principals, and principals received training in conducting rigorous and consistent observations from the education consulting firm, Cambridge Education. Ensuring a high degree of inter-rater reliability on the observation rubric was particularly important. It was critical that all peers, mentors, principals, and assistant principals had the same definitions for all four performance ratings in each domain’s components or teachers might perceive the system as unfair. More importantly, in an inconsistent system teachers might not receive appropriate feedback to improve their practice and student learning could languish. For peers and mentors, that meant reaching consistent classifications on four performance ratings in 17 different components. Cambridge Education used a methodical process to make sure everyone was measuring each component in the same way. Evaluators received at least 40 to 50 hours (and sometimes more) of training, including classroom instruction, one-on-one work with a trainer, shared sessions with multiple evaluators and a trainer, and reflection. Peers, principals, and mentors were not certified to conduct evaluations until they could demonstrate high inter-rater reliability. It was a training experience Brown called “intensive.”

By the beginning of the school year, including administrators, Cambridge Education certified 686 observers as “Accomplished” and ready to perform classroom observations. As of January 2011, Cambridge Education was still actively involved in calibrating the peer and mentor and principal evaluators, and the observation process had started in earnest. Nearly 10,000 observation cycles – pre-observation conference, observation, and post-observation conference – had been completed so far in the 2010-2011 academic year.

The Observation Cycle

In the former system, a tenured teacher only needed to be observed once every three years. The new evaluation called for a teacher to be observed three to eleven times throughout the year, depending on his or her needs. The observations were split between the peer or mentor and principal. The cycle, consisting of a pre-conference meeting, observation, and post-observation conference, was designed to encourage reflection. As Brown said, “One of the things that really appealed to us in Charlotte’s work was not only that it was rubric-based, but also that it had a very strong reflective piece of it for teachers because we believe a reflection is very powerful as teachers work to improve and get better at their craft.”

Before the pre-observation conference, a teacher completed a set of basic questions, including: What are your lesson objectives? How are the lesson objectives aligned with state curriculum standards? What data did you use to design this lesson? How did the data influence the planning of this lesson? How will you know if your lesson objectives were achieved? Peers and mentors then met with teachers to go over answers and ask further questions using a pre-conference guide document designed to stimulate greater reflection. After an observation was conducted, the

Multiple Measures of Teacher Effectiveness

completed rubric was loaded into a data management system where it could be viewed instantly by teachers, principals, and the peer or mentor teachers. Teachers had another chance to reflect on the lesson when they met with the observer in the post-observation conference.

While providing more precise information about teacher performance, the new observation system also established a platform for providing feedback and resources. Central to this was the adoption of the Lawson Talent Management System, an online performance management system that allowed peers, mentors, and principals to input observations, which were immediately available to teachers. The teacher, principal, or peer could then look at the assembled data and make decisions about appropriate follow-up professional development. In addition, HCPS was in the process of aligning its array of online and in-person professional development opportunities to each of the specific quadrants on the Danielson rubric. So if a teacher was rated as Developing on the Using Questioning and Discussion Techniques component of the Instruction domain, he or she would know exactly what trainings were available to address the areas of need.

Changing Role of the Principal

An outgrowth of the redesigned evaluations has been a changed role for principals, one that re-centers them as instructional leaders. Principals now spend more time in the classroom, part of an evaluation-mandated commitment to perform observations of every teacher, every year. Woodford explained, “A great deal of their time now is being spent doing observations because every teacher in every school, K-12, is to be observed by their principal at least one time. Principals are spending a great deal of time in this process, as in the past they haven't been required to do an observation annually of each teacher.” For some administrators, that has meant a shift in philosophy, and a challenge in organization. But, the shifting role only accentuated an accelerating trend in school administration. As McManus commented, the role of principal has evolved over the years into more of an “instructional leader,” rather than a “building manager type,” and the increased emphasis on observation and feedback in the new evaluation model “has pushed that area even more.”

The new evaluation also meant collaborating with peers and mentors. Peer teachers and principals discussed their observations — only after each has conducted his or her observation, so as not to bias an observation for either party — and the dialogue created a chance to provide teachers who needed help with focused support. For Woodford, the principals and peers were not isolated actors inputting information, but collaborators working together with the teacher to improve results. She explained, “Once the cycle’s complete, we’re all about helping teachers get better. The communication should be what this teacher needs in support. So, it’s important when a teacher’s struggling that the principal and the peer have conversations. That’s really started happening now.”

Principals were also being evaluated differently to reflect their new roles. Fifteen percent of a principal’s evaluation was now based on teachers’ feedback gathered in a 360-degree survey developed by Vanderbilt University. Another 40 percent was linked to student achievement — 30 percent based on overall student growth and 10 percent on the learning of high-needs students. A principal’s ability to retain his or her school’s most effective teachers was worth 5 percent on the new instrument. In addition, another 5 percent was based on the link between the principals’ ratings of a teacher and that teacher’s value-added measure. An exemplary evaluation for a teacher with less than average overall student growth would mean a lower evaluation rating for principals.

Value-added Measures

The incorporation of student achievement using a value-added measure was the third and final component of the new evaluation system, comprising 40 percent of a teacher’s evaluation. The

weight given to student achievement was actually less than the one previously used to allocate bonus money to HCPS teachers. Florida was a leader in incorporating assessments and accountability in its K-12 public education system. In 1973, it became the first state to require annual testing of nearly all students in certain grades and subjects.⁵ Five years later, Florida expected all high school students to pass an exit exam to earn a diploma. By the late 1990s and early 2000s, the state was experimenting with linking student test scores to teacher compensation. In these initial voluntary efforts, teachers could earn a five percent bonus if they could show significant achievement gains through a portfolio review process.⁶ HCPS and HCTA were willing partners in these early reforms around teacher evaluation and bonus-pay.

For the 2006-2007 school year, Florida developed its Merit Award Program (MAP), which gave districts money to pay top teachers bonuses. To participate, growth in student achievement needed to account for at least 60 percent in a district's determination of bonus-eligible teachers. HCPS signed on to MAP and designed its own system that incorporated student growth to allocate bonuses to teachers. However, unlike other districts that only included core subject teachers in MAP, HCPS had nearly all of its teachers participate in the new bonus initiative. HCPS's bonus-system combined a value added-like measure with principal evaluation to rank teachers by performance. Teachers in the top quartile or so received bonuses of around \$2,000. To implement its MAP program, HCPS worked together with HCTA to create over 50 different teacher comparison groups, pre- and post-tests for those subjects missing such measures, and a data management system capable of linking students to individual courses and teachers.

While the basic infrastructure needed to implement value-added measures was in place, the new student achievement component in the evaluation was more comprehensive than the one used in MAP. First, MAP only used student achievement to allocate teacher bonuses; growth was not incorporated into a teacher's overall evaluation. Another major weakness with the MAP system was it determined student learning growth based on single pre- and post-test instruments using only one year of data. More importantly, the measure did not adequately control for some school and student-level characteristics.⁷ The value-added measure in the new evaluation system promised to address these problems by integrating student achievement into 40 percent of a teacher's final evaluation, incorporating multiple pre- and post-assessments, and using a much more powerful model to calculate student achievement growth that incorporated three years of data. Finally, the percentile ranking was replaced by a scoring system which placed no artificial ceiling on excellence.

To develop its value-added measure, HCPS partnered with the University of Wisconsin's Value-Added Research Center (VARC) because of the group's extensive experience working with districts in Chicago, Milwaukee, and New York City. HCPS and VARC then worked on assessing the district's course testing map, a document that laid out all the pre- and post-assessments currently used in the district. Courses in need of improved assessments were identified and experienced teachers helped craft more and better pre- and post-tests. Once students were tested in a particular subject, the results were sent electronically to the district's data warehouse. There, the data were linked between students, their background characteristics, and the course teacher using a system of unique identification numbers. VARC then used the linked data to compute a value-added measure for every teacher in HCPS. The outcome was a score, ranging from positive to negative, that indicated how much growth a teacher's students had accomplished compared with that of an average district teacher for that particular subject. Positive value-added scores meant higher than average growth. Because 2010-2011 was the first year of implementation for the program, teachers' initial value-added scores included one year of data. Additional years of student achievement data would be incorporated until three years of data were accumulated. The value-added measure in the

Multiple Measures of Teacher Effectiveness

evaluation would then include the most recent three years in order to provide a more accurate and stable indicator of teacher effectiveness.

Communicating the New System

Communication about the new evaluation system has been key to winning support among teachers, community leaders and the general public. That drive for stakeholder support has included over 100 on-site presentations, which serve as opportunities to both communicate programmatic information and to receive feedback. As Brown explained, “It gives us great lift, tremendous lift, when they feel that they have input, and that they're being heard.” In addition, a six-hour course in the Danielson framework was available to all district personnel, as was a three-hour follow-up course. All schools received a copy of Charlotte Danielson’s “Frameworks for Teaching,” and all district personnel had access to an electronic copy. Pre-service professional development included an explanatory video narrated by teachers. Presentations to school faculty were conducted live and via podcast and an e-magazine on the new system was circulated. To promote community awareness of the initiative, Superintendent Elia and her staff met with a wide variety of stakeholders, including community groups, legislative bodies and university staff.

Moving Forward

The impact of a teacher evaluation model including multiple measures of teaching effectiveness was both concrete and intangible. A clear and objective observation instrument – the Danielson rubric – ensured that peer, mentor and principal evaluations were consistent, specific, and action-based. An immediate outcome of the observations and evaluations was targeted support for teachers to improve their practice. Combining student achievement growth with the peer and principal evaluations formed an equitable and quantifiable measurement of a teacher’s impact on his or her students. Finally, the multiple measures offered different points of view on a teacher’s performance, which many teachers perceived as being fairer than evaluations relying solely on the principal’s judgment or student achievement growth.

Teacher participation and leadership in designing and implementing the new evaluation system were key. As the stakeholders most responsible for student learning, teacher support and input were critical to the success of the initiative. Furthermore, communication with and continued training for all parties – teachers, principals, and peer and mentor evaluators – ensured the long-term viability of the evaluation model. The multiple measures system promised to give a more reliable and consistent evaluation of teacher performance, and provided a platform for future reforms. In the 2013-2014 school year, after three evaluation cycles under the new process had been completed, a new career ladder and compensation scale would go into effect, tied to the multiple-measure evaluations. Ultimately, teachers who were successful in cultivating student achievement would be rewarded, developing teachers would have actionable information to help them become stronger instructors, and student achievement would be centered as a focus of evaluation.

Figure 1: Percentage of HCPS Third Graders Scoring 3 or Above in Reading on the Florida Comprehensive Assessment Test

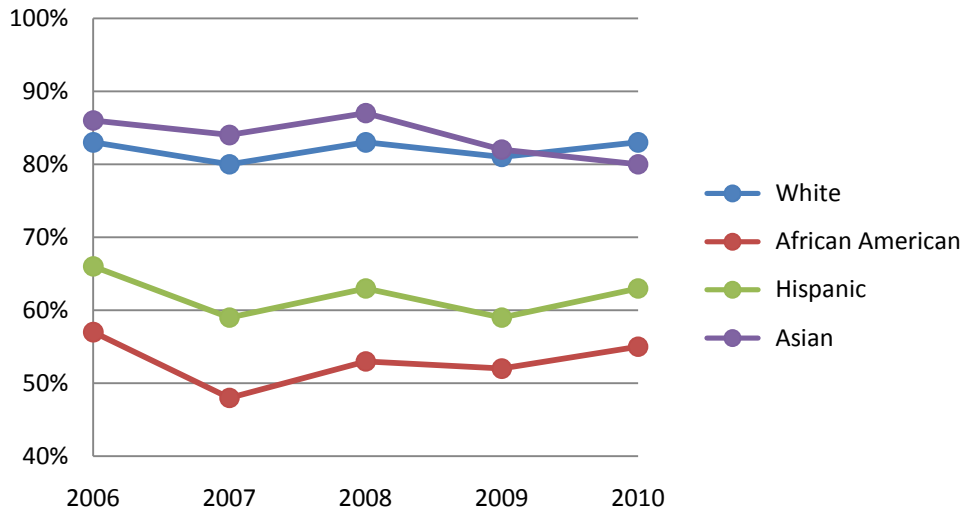
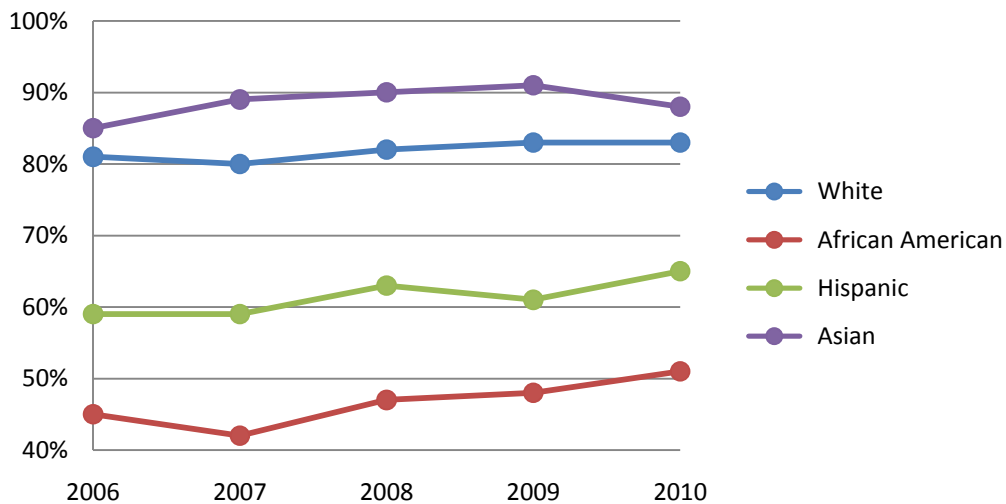


Figure 2: Percentage of HCPS Tenth Graders Scoring 3 or Above in Math on the Florida Comprehensive Assessment Test



Source: Florida Department of Education, "Florida Comprehensive Assessment Test Results." Retrieved April 2011 from <http://fcats.fldoe.org/>

Multiple Measures of Teacher Effectiveness

Figure 3: Excerpt from the Charlotte Danielson Observation Rubric Used by HCPS

					Performance rating					
					Requires Action <i>(0 points)</i>	Developing <i>(1 points)</i>	Accomplished <i>(2 points)</i>	Exemplary <i>(3 points)</i>		
Domain 2: The Classroom Environment					<i>(Domain % of evaluation score: 20%)</i>					
2a. Creating an Environment of Respect and Rapport (Component weight: 5%)	Classroom interactions, both between the teacher and students and among students, are negative, inappropriate, or insensitive to students' cultural backgrounds and are characterized by sarcasm, put-downs, or conflict.	Classroom interactions, both between the teacher and students and among students, are generally appropriate and free from conflict, but may be characterized by occasional displays of insensitivity or lack of responsiveness to cultural or developmental differences among students.	Classroom interactions between the teacher and students and among students are polite and respectful, reflecting general warmth and caring, and are appropriate to the cultural and developmental differences among groups of students.	Classroom interactions among the teacher and individual students are respectful, reflecting genuine warmth and caring and sensitivity to students' cultures and levels of development. Students themselves ensure high levels of civility among members of the class.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		
<i>Elements include:</i> Teacher interaction with students Student interactions with other students										
2b. Establishing a Culture for Learning (Component weight: 5%)	The classroom environment conveys a negative culture for learning, characterized by low teacher commitment to the subject, low expectations for student achievement, little or no student pride in work and no evidence that students believe that they can succeed if they work hard.	The teacher's attempt to create a culture for learning is partially successful, with little teacher commitment to the subject, little evidence that students believe they can succeed if they work hard, modest expectations for student achievement, and little student pride in work. Both teacher and students appear to be only "going through the motions."	The classroom culture is characterized by high expectations for most students, the belief that students can succeed if they work hard, and genuine commitment to the subject by both teacher and students, with students demonstrating pride in their work.	High levels of student energy and teacher passion for the subject create a culture of learning in which everyone shares a belief in the importance of the subject and the belief that students can succeed if they work hard. All students hold themselves to high standards of performance—for example, by initiating improvements to their work.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		
<i>Elements include:</i> Importance of the content Expectations for learning and achievement Student pride in work										

Source: Internal Hillsborough County Public Schools Document

Endnotes

¹ Daniel Weisberg, Susan Sexton, Jennifer Mulhern, & David Keeling, "The Widget Effect: Our National Failure to Acknowledge and Act on Teacher Differences." Brooklyn, NY: The New Teacher Project: 2009. Retrieved January 2011 from www.widgeteffect.org.

² Daniel Weisberg, Susan Sexton, Jennifer Mulhern, & David Keeling, "The Widget Effect: Our National Failure to Acknowledge and Act on Teacher Differences." Brooklyn, NY: The New Teacher Project: 2009. Retrieved January 2011 from www.widgeteffect.org.

³ Daniel Weisberg, Susan Sexton, Jennifer Mulhern, & David Keeling, "The Widget Effect: Our National Failure to Acknowledge and Act on Teacher Differences." Brooklyn, NY: The New Teacher Project: 2009. Retrieved January 2011 from www.widgeteffect.org.

⁴ Florida Department of Education, "Florida Comprehensive Assessment Test Results." Retrieved January 2011 from <https://app1.fldoe.org/FCATDemographics/Default.aspx>.

⁵ Herrington, C.D. & MacDonald, V.M. "Accountability as a school reform strategy: A 30-year perspective on Florida." In *Florida 2001: Educational Policy Alternatives*, (eds.) Carolyn D. Herrington & Katherine Kasten. Jacksonville: Florida Institute of Education, University of North Florida, 2001.

⁶ Hillsborough Classroom Teachers Association and School Board of Hillsborough County Contract, August 4, 2005. Retrieved January 2011 from <http://www.nctq.org/contracts/8-05.pdf>.

⁷ Letitia Stein, "Hillsborough's Merit Pay Experiment Benefits Affluent Schools," *St. Petersburg Times*, February 24, 2008. Retrieved January 2011 from http://www.sptimes.com/2008/02/24/Hillsborough/Hillsborough_s_merit_.shtml