Introduction

In an effort to advance K-12 education across the state of Louisiana, policymakers, educators and community leaders have put into action legislation designed to give administrators and teachers powerful accountability measures and tools that can be used to improve student achievement. The prior system of evaluation has been criticized as being too subjective and "[failing] to provide teachers and administrators with pertinent and objective data to understand and improve their effectiveness." Act 54 calls for the implementation of a two-fold evaluation mechanism, including both a value-added component based on student performance measures and a more traditional observation component based on a rubric of teacher and administrator actions. These models have been put into limited action on a pilot basis to determine the strengths and areas for improvement in these measures before the system is rolled out statewide in 2012-2013.

The purpose of this report is to communicate the process and outcomes of the 2011-2012 Compass Pilot to educators, policymakers, researchers, and members of the public.

Methodology

During the 2011-2012 school year, nine parish school districts and one Type 2 charter school participated in a pilot of the Louisiana's Comprehensive Evaluation Model (Compass). Compass involved several processes that were new to educators, and this pilot served as an opportunity for teachers and administrators to give feedback on each step in the process before full-scale implementation of the program. Beginning in 2011, teachers and administrators in the pilot schools received training on each aspect of the Compass teacher evaluation system. The following processes were introduced and assessed through the pilot. Each process is explained in greater detail in this section.

- Setting goals with teachers in Non-Tested Grades and Subjects (NTGS)
- Evaluating NTGS teachers' goals through the use of rubrics
- Determining the extent to which these teachers achieved their goals
- Rating teachers' classroom performance and administrators' school leadership skills on standardized rubrics
- Collecting standardized testing data and determining a value-added score for all teachers in testing grades and subjects
- Documenting, tracking, and communicating these rating to teachers and leaders using the Human Capital Information System (HCIS), a database of demographic information, value-added scores and rubric scores for each teacher

<u>Non-Tested Grades and Subjects:</u>

The initial phase of the pilot process required that teachers in Non-Tested Grades write quantitative goals, called Student Learning Targets (SLT), for their class. The rubric (which can be found in the Appendix) contained 2 elements: SLT Quality and Goal Attainment. SLT Quality was scored by the administrator at the beginning of the school year and measured the overall quality of the SLTs, specifically: the inclusion of an initial student assessment (baseline), a description of the indicators used to measure performance, and the alignment of the goals with current standards and GLEs. The Goal Attainment element of the rubric was scored at the end of the teaching period for each teacher and reflected the extent to which the teacher achieved the goals set at the beginning of the year. This score for each teacher

is reflected by the *NTGS Score* variable, and has a possible range of 1 to 5 as reflected by the average of each component on the attached rubric.

Rubrics and Observation-based Evaluation of Teachers:

Throughout the pilot school year, classroom teachers and administrators were evaluated through the use of rubrics (found in the Appendix). Teachers were evaluated on 11 performance standards clustered into the competency areas of Planning, Instruction, Environment, and Professionalism (see Table e). Trained administrators evaluated each teacher through a series of observation cycles, which included a pre-observation conference, an observation, and a post-observation conference. At each post-observation conference, administrator's notes and rubric scores were shared with the teacher, and strengths and areas for growth were discussed. For the purpose of statistical analysis, Planning has been assigned the label *Competency 1*; Instruction has been labeled *Competency 2*; Environment has been labeled *Competency 3*; and Professionalism has been labeled *Competency 4*. The values of Competency variables range from 1 to 5 and represent the average of the performance standards listed under each competency in Table e. The average of the scores received on all 11 performance standards is reflected in the *Performance Evaluation Score* variable.

Table e – Teacher Performance Standards listed by Competency

	Planning *Competency 1*	Instruction *Competency 2*	Environment *Competency 3*	Professionalism *Competency 4*
Performance Standards	PLANNING STANDARD 1: The teacher aligns unit and lesson plans with the established curriculum to meet annual achievement goals.	INSTRUCTION STANDARD 1: The teacher presents accurate and developmentally-appropriate content linked to real-life examples, prior knowledge, and other disciplines.	ENVIRONMENT STANDARD 1: The teacher implements routines, procedures, and structures that promote learning and individual responsibility.	PROFESSIONALISM STANDARD 1: The teacher engages in self-reflection and growth opportunities to support high levels of learning for all students.
	PLANNING STANDARD 2: The teacher designs lesson plans that are appropriately sequenced with content, activities, and resources that align with the lesson objective and support individual student needs.	INSTRUCTION STANDARD 2: The teacher uses a variety of effective instructional strategies, questioning techniques, and academic feedback that lead to mastery of learning objectives and develop students' thinking and problem-solving skills.	ENVIRONMENT STANDARD 2: The teacher creates a physical, intellectual, and emotional environment that promotes high academic expectations and stimulates positive, inclusive, and respectful interactions	PROFESSIONALISM STANDARD 2: The teacher collaborates and communicates effectively with families, colleagues, and the community to promote students' academic achievement and to accomplish the school's mission.
	PLANNING STANDARD 3: The teacher selects or designs rigorous and valid summative and formative assessments to analyze student results and guide instructional decisions.	INSTRUCTION STANDARD 3: The teacher delivers lessons that are appropriately structured and paced and includes learning activities that meet the needs of all students and lead to student mastery of objectives.	ENVIRONMENT STANDARD 3: The teacher creates opportunities for students, families, and others to support accomplishment of learning goals.	

Rubrics and Observation-based Evaluation of School Leaders:

Administrators were evaluated on 17 performance standards clustered into the competency areas of Ethics and Integrity, Instructional Leadership, Strategic Thinking, Resource Management, and Educational Advocacy (see Table f). Evaluators from outside of the school conducted an interview, a teacher and staff survey, and a site visitation. After the site visitation, evaluators participated in conferences with the school leader in order to share comments and scores from the site visitation rubric and to make suggestions for improvement. For the purpose of statistical analysis, Ethics and Integrity has been assigned the label *Competency 1*; Instructional Leadership has been labeled *Competency 2*; Strategic

Thinking has been labeled *Competency 3*; and Resource Management has been labeled *Competency 4*; Education Advocacy has been labeled *Competency 5*. The values of Competency variables range from 1 to 5 and represent the average of the performance standards listed under each competency in Table f. The averages of the scores received by school leaders on this rubric are, again, reflected in the *Performance Evaluation Score* variable.

Table f – Administrator Performance Standards listed by Competency

	Ethics and Integrity *Competency 1*	Instructional Leadership *Competency 2*	Strategic Thinking *Competency 3*	Resource Management *Competency 4*	Educational Advocacy *Competency 5*
Performance Standards	STANDARD 1: The leader demonstrates compliance with all legal and ethical requirements.	STANDARD 1: The leader establishes goals and instructional and leadership expectations.	STANDARD 1: The leader engages stakeholders in determining and implementing a shared vision, mission, and goals that are focused on improved student learning; are specific, measurable, achievable, relevant, and timely (SMART); and that anchor plans for school improvement.	STANDARD 1: The leader manages time, procedures, and policies to maximize instructional time as well as time for professional development opportunities that are aligned with the school's goals.	STANDARD 1: The leader provides opportunities for multiple stakeholder perspectives to be voiced for the purpose of strengthening school programs and services.
	STANDARD 2: The leader publicly articulates a personal educational philosophy or set of beliefs to coworkers.	STANDARD 2: The leader plans, coordinates, and evaluates teaching and the curriculum.	STANDARD 2: The leader formulates and implements a school improvement plan to increase student achievement that is aligned with the school's vision, mission and goals; is based upon data; and incorporates research-based strategies and action and monitoring steps.	STANDARD 2: The leader allocates financial resources to ensure successful teaching and learning.	STANDARD 2: The leader stays informed about research findings, emerging trends, and initiatives in education in order to improve leadership practices.
	STANDARD 3: The leader creates a culture of trust by interacting in an honest and respectful manner with all stakeholders.	STANDARD 3: The leader promotes and participates in teacher learning and development.	STANDARD 3: The leader analyzes data from student results and adult implementation indicators to monitor the impact of the school-wide strategies on student learning.	STANDARD 3: The leader creates a safe, healthy environment to ensure effective teaching and learning.	STANDARD 3: The leader acts to influence national, state, and district and school policies, practices, and decisions that impact student learning.
	STANDARD 4: The leader models respect for diversity.	STANDARD 4: The leader creates a school environment that develops and nurtures teacher collaboration.			

In addition to the data on teacher and administrator performance collected by objective evaluators using rubrics, teachers and administrators were asked to complete reflective self-evaluations with identical rubrics. The average of the scores of each performance standard on this rubric is reflected in the *Self Evaluation* variable.

Value-Added Measures:

At the end of the Compass pilot each teacher in a tested-grade or subject received a value-added score based on student performance measures. This score was designed to reflect the impact that just the teacher had on the student's academic performance and was obtained by controlling for factors that influence student achievement such as prior years' test scores, free/reduced-lunch status, attendance, and disciplinary issues. While this report is not primarily concerned with the validity or efficacy of the value-added model, there have been many studies that have provided evidence that these measures are appropriate and effective for determining the impact of individual educators on student performance (Bock & Wolfe, 1996; Sanders & Rivers, 1996; Gordon, Kane, and Staiger 2006; Hanushek 2009;

Chetty, Friedman, Rockoff, 2011). Evidence from Louisiana's pilot of the value-added model has demonstrated this model to be very reliable from year to year particularly when compared with others from around the country (see Appendix). The *VAM* variable found in the current study is a conversion of teachers raw value-added scores to a scale with a range from 1 to 5 measured to one decimal place based on the range of the actual value-added score. This conversion eliminates negative value-added scores that would invalidate certain statistical methods.

Qualitative Data Collection:

In addition to the collection of this quantitative data, a qualitative study was conducted using surveys, interviews, and focus groups of teachers and administrators from the Compass pilot. The survey included 542 total participants, including 438 teachers (80.8%), 77 principals (14.2%), and 27 central office staff (5.0%). The survey, which contained both multiple-choice and open-ended questions, assessed participants' perceptions of the Compass system overall, the Value-Added Model, Student Learning Targets, the clarity and relevance of rubric items, and the ease of implementation of the Compass model. Structured interviews were used to explore in more detail participants understanding and satisfaction with the performance standards and competencies of the Compass rubrics. The focus groups utilized a standardized open-ended protocol to gain insight into participants' understanding of the Compass system, acceptance of the Compass rubric, value-added measures, and SLT system for non-tested grades and subjects.[1]

Results

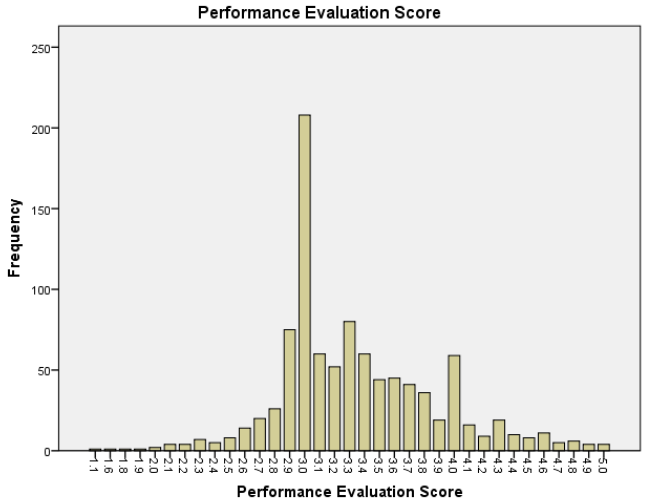
This paper will report and analyze the quantitative and qualitative data collected from multiple sources through the Compass pilot. Observational data was gathered through the use of rubrics reflecting teachers' and school leaders' skill in predefined competency areas. Teacher effectiveness was also measured by way of a value-added model (VAM) and scores from both evaluation systems were paired to allow for comparison and statistical analysis. Throughout the process surveys, interviews and focus groups were conducted in order to better understand how participants were interacting with the tools and methods introduced by the pilot and to gather participants' feedback on possible improvements to the systems.

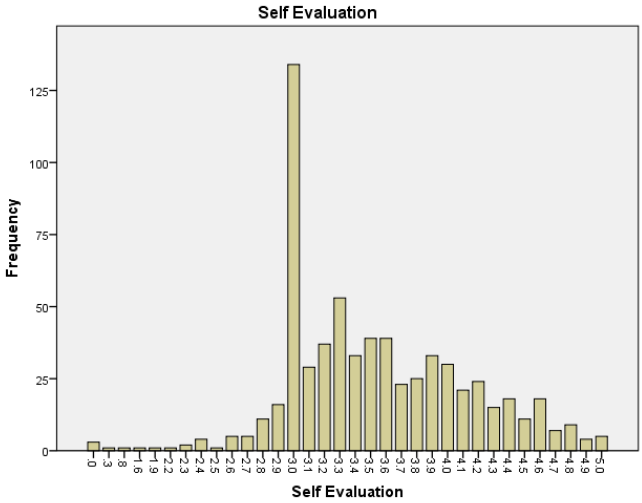
Quantitative Analysis

The next section will report the quantitative results from statistical analysis of the data produced by the Compass pilot. In the first subsection, descriptive measures will be used to give evidence of the reliability of the scoring rubrics and to explore any irregularities in the data. In the second subsection, inferential statistics will be applied to determine whether correlation is present in the data and to explore the implications of any correlations that are found.

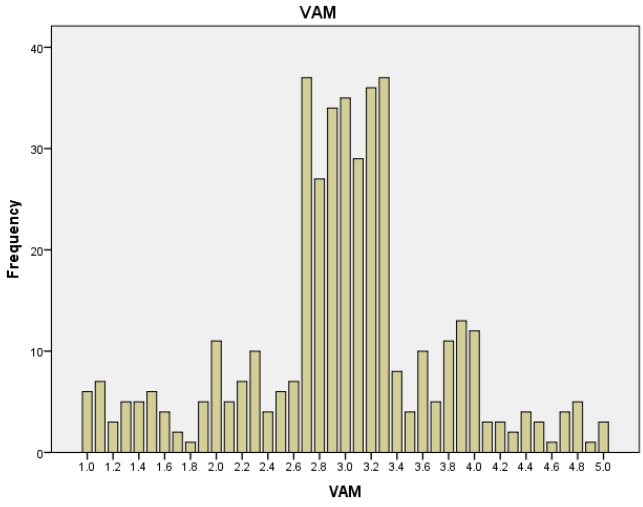
[1] Compass Pilot Feedback Summary Report, 2012, p. 3

Descriptive Statistics:





The distribution of the *Performance Evaluation Score* and the *Self Evaluation* scores are negatively skewed. This is consistent with expectations regarding distributions of scores of observational evaluations, particularly self-reported evaluations. Even given the skewness of the distributions, the distribution demonstrates, as expected, that among the population of teachers many are average or close to average with fewer and fewer performing far below or far above their collegues. However, there is a clustering at the score of 3 (the mode) for both the *Performance Evaluation Score* and the *Self Evaluation* score. This mode accounts for 21.6 % of the total recorded scores in the *Performance Evaluation* variable and 20.4% of the total recorded scores for the *Self Evaluation* variable.



The *VAM* variable is not as evenly distrubeted as the metrics above. This distribution is not normally distributed and is multi-modal. There is a high frequency of scores within a very tight range; 55.8% of the scores fall between 2.7 and 3.3.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Performance Evaluation Score	965	1.1	5.0	3.340	.5354
Self Evaluation	658	.0	5.0	3.517	.6263
VAM	421	1.0	5.0	2.967	.7959
NTGS Score	181	1.7	5.0	3.103	.6314
Valid N (listwise)	1				

Figure c

The minimum, maximum, mean, and standard deviation of each variable are reported in Figure c. The variables with means greater than 3 (the middle possible score on each rubric) are *Performance Evaluation Score* (3.340), *Self Evaluation* (3.517), the *NTGS Score* (3.103). The mean of the *VAM* variable (2.967) is slightly lower, but very close to the middle possible score. There appears to be a slight difference between ratings based on statistical student growth metrics and those based on evaluator judgements; measures based on evaluator judgements are slightly higher overall than value-added measures. While the performance evaluations and the value added models are meant to be complementary metrics for teacher effectiveness, the fact that these variables have different distributions is not particularlly alarming as they measure effectiveness in very different ways – performance evaluations measure the visible signs of good teaching, while the value-added model measures student achievement as a result of teaching. The only conclusion that might be drawn based on these distributions are that raters found educators (and educators found themselves) to be somewhere between "Proficient" and "Accomplished" most of the time.

Correlations

		Performance Evaluation Score	VAM	Self Evaluation	NTGS Score
Performance Evaluation Score	Pearson Correlation	1	.226**	.372**	.444**
	Sig. (2-tailed)		.000	.000	.000
	N	965	316	592	158
VAM	Pearson Correlation	.226**	1	.184**	.[b]
	Sig. (2-tailed)	.000		.008	.000
	N	316	421	206	3
Self Evaluation	Pearson Correlation	.372**	.184**	1	.240**
	Sig. (2-tailed)	.000	.008		.009
	N	592	206	659	116
NTGS Score	Pearson Correlation	.444**	.[b]	.240**	1
	Sig. (2-tailed)	.000	.000	.009	
	N	158	3	116	182

**. Correlation is significant at the 0.01 level (2-tailed).

b. Cannot be computed because at least one of the variables is constant.

<u>Inferential Statistics:</u>

According to the correlation table (Figure a), the *Performance Evaluation Score*, the score obtained from Compass rubrics, shows a weak positive correlation with the *VAM*, the converted value-added score obtained by teachers, with a Pearson r of .226 and an r-squared of .051. The *Performance Evaluation Score* is also moderately correlated with the *Self Evaluation* score (Pearson r = .372) and the *NTGS Score* (Pearson r = .444), lending some evidence as to the reliability of this rubric.

Correlations

		Performance Evaluation Score	Competency1	Competency2	Competency3	Competency4
Performance Evaluation Score	Pearson Correlation	1	.897**	.906**	.918**	.863**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	965	965	965	965	965
Competency1	Pearson Correlation	.897**	1	.778**	.759**	.696**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	965	965	965	965	965
Competency2	Pearson Correlation	.906**	.778**	1	.793**	.682**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	965	965	965	965	965
Competency3	Pearson Correlation	.918**	.759**	.793**	1	.733**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	965	965	965	965	965
Competency4	Pearson Correlation	.863**	.696**	.682**	.733**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	965	965	965	965	965

**. Correlation is significant at the 0.01 level (2-tailed).

According to the correlation table (Figure b), the *Performance Evaluation Score* is highly correlated with the scores on each competency: *Competency1* – r = .897, *Competency2* – r = .906, *Competency3* – r = .918, *Competency4* – r = .863. Additionally there is a high level of correlation between the different competencies, again evidence of internal consistency and reliability. *Competency1* is highly correlated with *Competency2* (r = .778) and *Competency3* (r = .759) and moderately correlated with *Competency4* (r = .696). *Competency2* is highly correlated with *Competency3* (r = .793) and moderately correlated with *Competency4* (r = .682). Finally, *Competency3* and *Competency4* are highly correlated (r = .733).

Correlations

		VAM	Competency1	Competency2	Competency3	Competency4
VAM	Pearson Correlation	1	.216**	.245**	.240**	.125*
	Sig. (2-tailed)		.000	.000	.000	.027
	N	421	316	316	316	316

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

According to the *VAM* correlation table (Figure d), the *VAM* score is most highly correlated with *Competency2* (r = .245), *Competency3* (r = .240), *Competency1* (r = .216), and is least correlated with *Competency4* (r = .125).

Qualitative Analysis

This section will summarize the results and conclusions of a previous qualitative study focused on validating the Teacher Competencies and Performance Standards components of Compass. These results have already been reported and many of the solutions proposed have already been incorporated in revisions to Compass. While these concerns are reflective of a much earlier point in implementation, they still deserve rehashing in order to demonstrate the learning

that occurred as a product of the Compass pilot, as well as, to demonstrate the strides made since the initiation of the project. According to that study, responses centered around four main concerns:

> "There were concerns about design (e.g., why formal observation is more effective than informal observation), implementation (e.g., rushed timelines, confusing requirements), equity (e.g., evaluator's observations are reliable across schools/districts), and burden (e.g., too much paperwork and excessive time demands). Overall, there is considerable uncertainty about the system."[2]

Concerns about Value-Added:

Issues and concerns with design of the system were most frequently associated with the Value-Added Model; there were parts of that method which were at that time uncomfortable to teachers and administrators. Teachers in the interviews were skeptical that standardized test achievement could be used to measure teacher effectiveness. According to the survey, half of respondents felt "confident" or "prepared" for the implementation of the Value-Added Model; however, far less (only around 30%) felt that value-added measures would provide meaningful information. In the context of Compass, this means that more effort was needed to communicate the foundation for the validity of the Value-Added component of the teacher evaluation system. This effort would have to include reasoning with appeal to the population more broadly, since the distrust of the system stemmed simply from a lack of understanding.

Concerns about Implementation:

Implementation was a challenge in the process of completing the Compass Pilot. Respondents mentioned the difficulty of completing the five-point rubric because the levels of achievement were difficult to differentiate. Some teachers even thought the top level of achievement was unattainable by many teachers. These problems were compounded by what teachers categorized as insufficient training. Many teachers were confused about the requirements and deadlines, and felt they were still (at the time of the survey) trying to figure out what to do. The final challenge to implementation was the noted difficulty of transferring between the paper-based evaluation system and the HCIS electronic platform. Many respondents pointed out glitches in the electronic system and other settings that made the platform difficult to navigate or frustrating to use for long periods of time.

Concerns about Equity:

Many teachers and school leaders voiced concerns about equity; that is, how fairly the system will be implemented across the diverse districts and schools in Louisiana. Cheating or "gaming the system" by forging artifacts was a concern with some; however, the report makes it clear that there has been no evidence of this happening. Other concerns with equity stemmed from the application of the standards to differing grade levels or subjects which could lead to different interpretations of the standards. Ultimately, it was noted by many that the quality and the credibility of the evaluation process was only as good as the evaluators implementing it, and that quality could vary depending on the individual evaluators or the prior relationships between the observer and the observed.

[2] Compass Pilot Feedback Summary Report, 2012, p. 18

Concerns about a “Burdensome” amount of Requirements:

The last major area for concern listed in the validation study was the burden of the Compass requirements. According to respondents the two most burdensome aspects of the process were the documentation of the evaluation process and the evidence needed of performance. Specifically, the lengths of the documenting paperwork (13 pages and 17 pages for teachers and principals respectively) were mentioned by respondents. Certainly a part of this perceived burden came from an actual overabundance of paperwork and an onerous amount of check-in steps throughout the process; however, the validation study also mentions participant misunderstandings as a factor which contributed to the sense of being overwhelmed. For instance, a misunderstanding of the role of teacher documentation led certain teachers to over emphasize this aspect of the process, dedicating valuable hours to something that was non-essential. Some aspect of the burden associated with the Compass pilot is likely an associated effect of the challenge of implementation of such an ambitious task in such a short timeframe.

While the concerns reported above focused solely on areas for improvement within the Compass system, there were considerable positive reactions reported by the study. Some in particular that bear mention