Balasubramanian, N., & Wilson, B. G. (2011). Continuous Improvement for K12 School Systems: An Assessment Tool for Educators. HarnessData White Paper.

Continuous Improvement for K12 School Systems: An Assessment Tool for Educators

Abstract

This paper introduces, HarnessData®, a web-based application for educators to analyze assessment data longitudinally to validate instructional effectiveness. The application has three powerful diagnostic tools to visually represent test performance of groups of students based on two factors: criterion-referenced achievement and value-added growth. First, a quadrant model contributes two necessary and essential questions for effective data inquiry using professional learning communities. Second, the strength charts present data on student growth, achievement and performance on state content standards and sub-content areas. Finally, the evidence of learning links student performance on state tests to district diagnostic and formative assessments across all subjects. By identifying gaps and mastery in learning across classrooms, the three tools could help improve and sustain system-wide continuous improvement practices.

Keywords

Accountability; data analysis; evaluation; instructional practices; school/teacher effectiveness

Just as students have a goal of reaching and exceeding standards of achievement, our teachers should excel on a standard of growth and achievement to help all students be fully prepared for college and their careers. For over four decades (Millman, 1997), one large urban public school system, the Dallas Independent School District, has examined teacher effectiveness using a variety of tests, including norm-referenced, criterion-referenced and state tests to engender system-wide continuous improvement practices. Their extensive research has highlighted significant differences in student achievement based on instructional effectiveness (Babu & Mendro, 2003). Yet information on the characteristics of effective pedagogy and everyday instructional practices of good teachers continues to be elusive and "decidedly meager" (Popham, 1997). During a recent press release on their 10-year $40 million initiative to guarantee the effectiveness of teachers, a reporter asked the panelists from the higher education institutions “How are you going to measure teacher effectiveness?” A panelist from Winona State University candidly replied “One of the important things about this incredible network and collaboration is none of us can pinpoint the answer to that question. What we can say is that we will work together to invent new models of demonstrating the effectiveness of our teachers” (Bush Foundation, 2009).

In this article, we introduce a web-based application for school administrators and teachers to analyze assessment data longitudinally and estimate instructional and program effectiveness. The application, called HarnessData® (and GME2), has three powerful diagnostic and continuous improvement tools in its current iteration to identify high-level patterns of performance with extensive drill-down capabilities. GME2™ is an acronym for Growth Model for Educational Excellence. Information is power. Educators can drill-down by district, school, department, teacher and subgroups based on their access level. They can easily compare schools, departments, teachers, and subgroups within the district.

The first tool in HarnessData®, called "PLC Quadrants," presents a graphic, growth-by-achievement-performance summary using annual standardized test data (Balasubramanian & Bankes, 2009; Balasubramanian & Muth, 2010). Achievement is depicted on the y-axis and growth is depicted on the x-axis. Using the four-quadrant model, the "PLC Quadrants" graphically depicts how much progress each student has made in one year regardless of their incoming achievement level. The achievement level of students in the quadrant model is criterion-referenced in that it illustrates how students stack up against established benchmarks of achievement. The growth in the model is value-added in that, to compute individual students' annual gains, it focuses on their year-to-year improvement by quantifying each students' current year achievement over their own prior year achievement on a common scale. The common scale in our model uses a test-independent measure of student achievement in a standardized test called Performance Index (this is discussed in the next section).

The second tool in HarnessData®, "Strength Charts," presents the performance mastery and needs of individual students across the different content standards and sub-content areas in reading, writing, mathematics and science. With these Strength Charts, users have the advantage of examining individual student performance relative to their own past performance (value-added), relative to their peers (normative), and relative to the standards (criterion-referenced). The third tool in HarnessData®, ongoing "Evidence of Learning," helps with tracking the progress of students across the different quadrants in the school/district diagnostic and formative assessments throughout the school year. The powerful HarnessData® diagnostic and continuous improvement tools provide constructive and timely feedback to students and educators to track individual student, department, school, and district progress on a variety of summative and formative assessments, even from the subjects not tested in state annual assessments.

The seven guiding principles underlying the development of our application and the associated methods are: (a) diagnostic tools are essential to further continuous improvement practices; (b) organizational goals should guide how this tool is used and not the other way round; (c) respect for individual students and educators; (d) tools are guided by one's values and users must guard against any unintended (ab)use (Black, 2001); (e) use of the application needs to be fully integrated into the values of the organization and fit into the lives of all stakeholders; (f) graphic representations are more accessible than tables, text, or the more usual bar charts and boxplots; and (g) providing growth data relative to initial levels of achievement is more useful than information provided by post-assessments alone.

The roadmap we present below uses the five W's and one H maxim from *The Elephant Child* (Kipling, 1912): "I kept six honest serving men. They taught me all I knew. Their names were *What* and *Why* and *When* and *How* and *Where* and *Who*." To highlight the tools' utility for driving continuous improvement in K-12 systems using credible evidence, we apply these five W's and one H framework on two concurrent projects using simulated data. The first phase of the project, “The Who→What→How Process,” provides good data and credible evidence (Mathison, 2009), is cost-effective to implement. The second phase of the project, “The Where→When→Why Process,” is somewhat less articulated because of the nature of the questions being asked. The second phase is more labor intensive and more culturally dependent on the values and intentions of local participants.

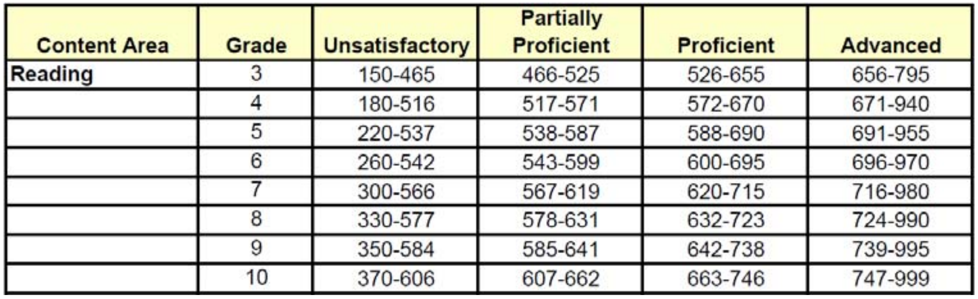
Making of Meaning of Scalable Scores in Standard Assessments

Before delving into the first phase with the "who," we should understand criterion-referenced standardized tests and how they are scored. The primary purpose of standardized tests is to obtain an annual snapshot measure of individual student learning and performance. Since the authors are from Colorado, we explain the concepts in this article using the Colorado Student Assessment Program (CSAP). The first author has evaluated statewide performance standards over several years and served as a content expert in Assessment Committees that reviewed the Content, Alignment, Validity and Cut Scores (for Standard Setting) for Science. Federal law has required all 50 states to embrace standards-based educational reform by developing aligned systems of state content standards, assessments, and performance cut-points (Barton, 2009). The common scale we use in our model can be applied to the scale scores on annual federally mandated state tests. A scale score is a transformation of a raw score (number of items answered correctly by students) into an equal-interval scale, using cut scores determined through the process of standard setting.

Large-scale assessments are constructed by including a "sample" of items from the different content standards and sub-content areas. For example, there are four content standards for the CSAP Reading test: reading comprehension, thinking skills, use of literary information, and literature. The four sub-content areas for the CSAP Reading test are fiction, nonfiction, vocabulary, and poetry. Students are tested on a timed standardized assessment in reading, writing, mathematics and science. They obtain a raw score based on their performance on all the scored items in the tests. Meanwhile, after scoring all the tests, the assessment contractor for the state (CTB/McGraw-Hill) presents an ordered item booklet to participants at the "Bookmark Standard Setting" meeting. The booklet has all the items rank ordered from the easiest item to the hardest item. This ordering also has Item Response Theory (IRT) scale score locations beside each item. These scale score locations are based on the ability levels of students. At any given location, students with the knowledge, skills and abilities at that level will have at least a 67% chance of responding successfully to that item. After two full days of discussions on what each item measures, why each item is more difficult than the items that precede it in the booklet, the participants as content experts in this collaborative enterprise then set "cut points" for what students at the different performance levels should know and be able to do. The standard setting is a rigorous process spanning several days and several rounds of placing and negotiating cut points to reach consensus through extended discussions and reflecting on the impact data. As Lewis, Green and Mitzel (1998) report, standard setting "is a complex process involving educational, psychological, statistical, and ultimately, political considerations" (p. 9). After standard setting, student raw scores are reported in terms of scale scores and performance levels.

In Colorado, student scores are reported using four performance levels—unsatisfactory, partially proficient, proficient and advanced. The range of possible scale scores varies by grade and subject across the four performance levels. For illustrative purposes, the highest obtainable scale score (HOSS) and the lowest obtainable scale score (LOSS) across the performance levels for the different grades in reading are provided in Figure 1. Similar tables exist for Spanish Reading (for grades 3 & 4), Writing (for grades 3 to 10), Spanish Writing (for grades 3 & 4), Mathematics (for grades 3 to 10) and Science (for grades 5, 8, & 10) in Colorado (CDE, 2008).

Suppose you are a teacher or parent of a fourth grader whose score increased from 530 to 560 in reading. You want to know how this individual is doing. What do you think? At first glance, it appears that the student has increased 30 points and that seems like it was an improvement. To understand how to interpret this seeming increase, you must know that CSAP uses a common scale. Student scores from the third grade reading and fourth grade reading tests, like the rest of the assessments are linked statistically to create a vertical scale. The vertical scale for CSAP reading is based on a "common item nonequivalent groups linking design" (Briggs & Weeks, 2009). So is 30 points increase an improvement?



Content Area	Grade	Unsatisfactory	Partially Proficient	Proficient	Advanced
Reading	3	150-465	466-525	526-655	656-795
	4	180-516	517-571	572-670	671-940
	5	220-537	538-587	588-690	691-955
	6	260-542	543-599	600-695	696-970
	7	300-566	567-619	620-715	716-980
	8	330-577	578-631	632-723	724-990
	9	350-584	585-641	642-738	739-995
	10	370-606	607-662	663-746	747-999

Figure 1. Performance level scale ranges for CSAP Reading (CDE, 2008).

On examining the first row of Figure 1, you see that the student scoring 530 in third grade reading is greater than the LOSS for proficient (526). This means the student was proficient in third grade. When this student scored 560 in the fourth grade reading test, however, it is less than the HOSS for partially proficient (571). The student actually went down a performance level! So with a 30 point increase, the student did not do better this year but actually worse than the prior year. How can teachers, administrators and parents make meaning of all this?

To help educators better understand scale scores and make them "intuitively comprehensible" (Popham, 1997), we have created a simple measure to estimate the extent of proficiency within a performance level and added it to the performance level and called it Performance Index (PI). The PI is a measure of student achievement in a standardized test (Perie, Weiss, Kurtz, & Dunn, 2009). PI is calculated by performing linear transformations of students' scale scores on vertically linked scale scores in annually mandated state tests. If tests are not linked with vertical scales, we cannot compare the performance of students from one grade to the next because different skills might be assessed. For example, in third grade CSAP reading test, students should demonstrate what they know and are able to do in one content standards and three sub-content areas — reading comprehension, fiction & poetry, nonfiction and vocabulary. In the fourth grade CSAP reading test, students should demonstrate what they know and are able to do in three additional content standards: thinking skills, use of literary information and literature.

To illustrate the logic of the formula, we return to our example above. Just like student grades are assigned numbers: D = 1.00, C = 2.00, B = 3.00 and A = 4.00, we have assigned numbers to the four *performance levels*: unsatisfactory = 1.00, partially proficient = 2.00, proficient = 3.00 and advanced = 4.00. To estimate the *extent of proficiency* within a performance level, we will normalize the scale score (split into 99 equal increments) with the following linear transformation to compute the incremental proficiency.

Incremental Proficiency = (Student Scale Score - LOSS)/(HOSS - LOSS)

Since some students' scores can be close to the LOSS and HOSS, to keep them between 0.01 and 0.99, we add one to the denominator. The formula now becomes

Incremental Proficiency = (Student Scale Score - LOSS)/(HOSS - LOSS + 1)

Performance Index (PI) = Performance Level (PL) + Incremental Proficiency (IP)

Applying this formula to the scale score 530, we can see from Figure 1 that the performance level is 3.00 (proficient). The Incremental Proficiency = (530 - 526)/(655 - 526 + 1) = 0.03. Therefore, the Performance Index of a student with a scale score of 530 in the third grade reading test is 3.03.

Similarly, applying the formula to the scale score of 560, we can see from Figure 1 that the performance level is 2.00 (partially proficient). The Incremental Proficiency = (560 - 517)/(571 - 517 + 1) = 0.78. Therefore, the Performance Index of a student with a scale score of 530 in the fourth grade reading test is 2.78.

Knowing students' PI is helpful because if we know their PI in two consecutive years, their value added growth from one year to the next is simply the change in PI.

Value-Added Growth = Change in PI = Current Year PI - Prior Year PI.

Applying this formula to our example with the student increasing scale score from 530 in third grade reading to 560 in fourth grade reading, we see that the value-added growth is 2.78 - 3.03 = -0.25. The negative sign shows that after one year of instruction, the teacher and school have not added value to this student but actually seen the student to decline by a quarter of an incremental proficiency. A whole host of contextual variables might have contributed toward this decline, including curricular alignment, school culture, teacher effect, student attendance, student behavior, and home environment. More on the contextual variables later.

HarnessData®, our web-based application, computes these performance indices for each student for each subject and the individual content standards and sub-content areas (see Figs 2, 3 & 4). HarnessData® is written in Microsoft® ASP.net 3.5 with SQL Server 2005 on Windows 2008. Unlike traditional data analysis with an oversimplified approach that relies on tracking "group means," the three tools in HarnessData® support continuous improvement goals of schools and districts by facilitating a more tailored approach to interpreting assessment data for individualized instruction. Our proprietary tool can be accessed by anyone interested at https://HarnessData.org with a username and password. Please e-mail the first author to request a username and password. We now address where we want educators to focus their attention, which is classroom instruction and instructional effectiveness.

Phase 1: Using PLC Quadrants and Strength Charts

To share and leverage the successes of students, teachers, and schools, the first phase starts with the "who," then delves into the "what," and concludes with the "how." We describe in this section the tools that help us with this first phase of inquiry. The scatterplot with these two dimensions—growth on the x-axis and achievement along the y-axis—produces four quadrants (Figure 2):

- High-achieving students doing better this year than last year [NE Quadrant];
- High-achieving students doing worse this year than last year [NW Quadrant];
- Low-achieving students doing better this year than last year [SE Quadrant].
- Low-achieving students doing worse this year than last year [SW Quadrant];

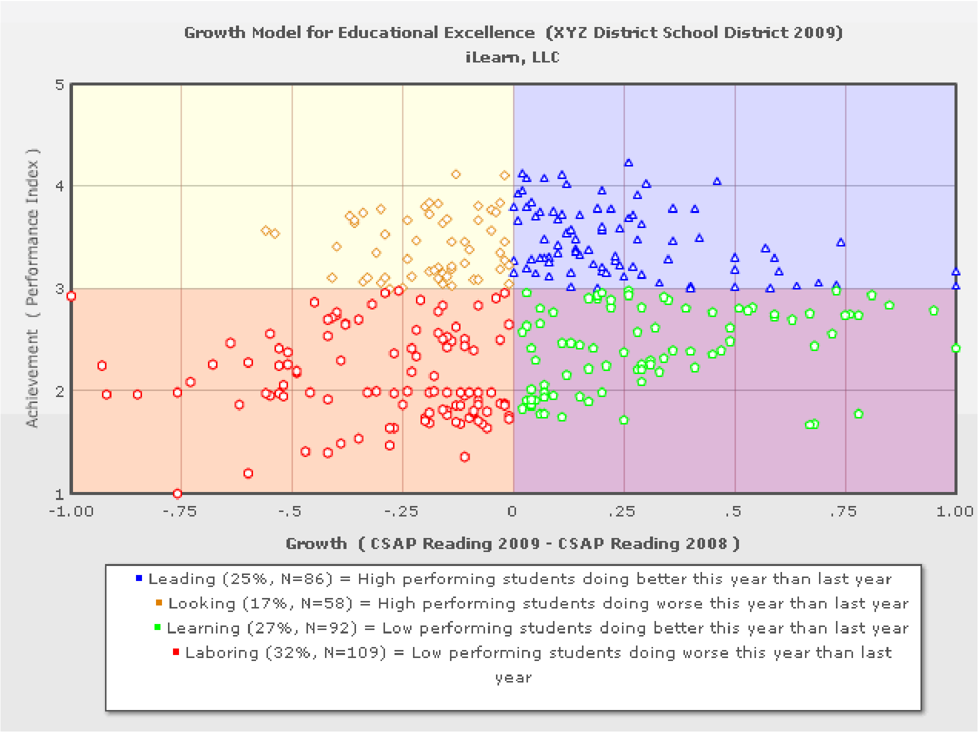


Figure 2. Quadrant model, displaying the growth and achievement of individual students.

A prominent educational author and consultant, Doug Reeves, uses the terms leading, learning, losing, & lucky in his leadership for learning framework (2006) using "antecedents of excellence" and "achievement of results" to describe schools. We have adapted these terms to make them explicit to educators and refer to the NE quadrant in our model as the "leading" PLC quadrant, the NW quadrant as the "looking" PLC quadrant instead of what Reeves calls "lucky," the SW quadrant as the "laboring" PLC quadrant instead of what Reeves calls "losing," and the SE quadrant as the "learning" PLC quadrant.

The horizontal axis is set at 3.00 (proficient) and the vertical axis is set at 0.00. The CSAP tests tend to measure most accurately near the cut scores (CTB McGraw-Hill, 2009), which are 2.00, 3.00 and 4.00 in our model. A student with a PI ≥ 3.01 is considered high achieving. A student with a PI ≤ 2.99 is considered low achieving. The standard error of measurement on growth with change-in-PI averages to zero with large sample sizes. Therefore, the criterion for the value-added growth is zero and it is an empirically defined quantity. A student with change-in-PI ≥ 0.01 is doing better this year. A student with change-in-PI ≤ minus 0.01 is doing worse this year.

To track instructional effectiveness more precisely (Figure 3) while accounting for contextual factors that might impact student achievement, administrators can analyze student growth and achievement information further by subgroups (gender, ethnicity, talented and gifted students, students with disabilities, economically disadvantaged students, students with limited English proficiency) and students who have been continuously enrolled at the school/district. Positive and negative trends/patterns in student performance across these subgroups can be observed by school, department, grade level and teacher(s).

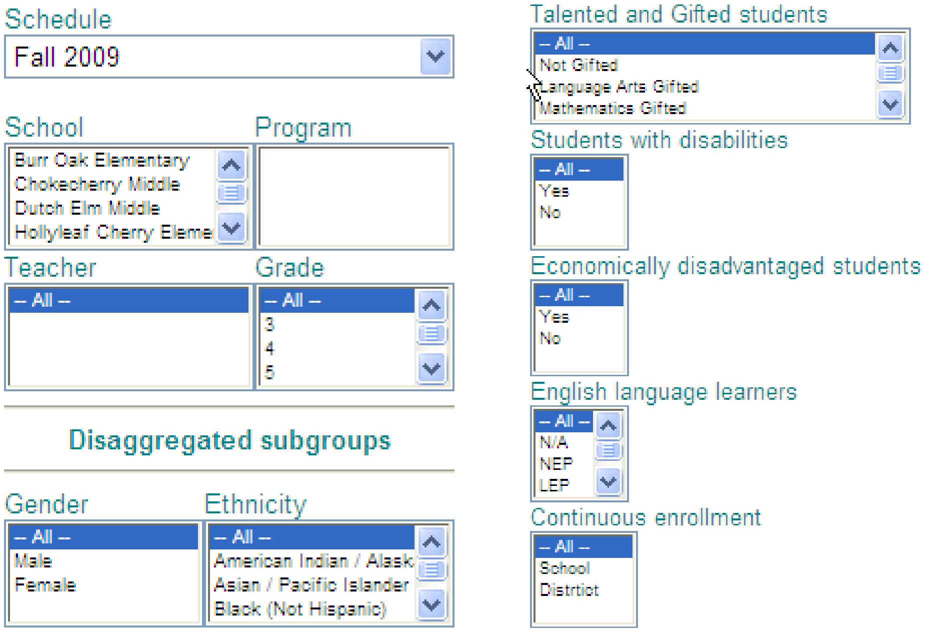


Figure 3. Contextual variables impacting student achievement.

The Who→What→How Questions

The tools described above provide data for continuous improvement inquiry. As researchers, we begin by asking:

- *Who* is making a difference (based on evidence observed year-to-year on student achievement *and* growth data on standardized tests represented by the quadrant model)?

Individual teachers and programs are identified that seem to be making substantial positive impact on student learning. These then become foci for further inquiry and sharing.

With the help of administrators who know these teachers, we then ask:

- *What* effective practices are the exemplary teachers using?
- *What* resources, methods, and interventions do they frequently utilize?
- *What* instructional strategies work with specific subgroups or populations of students?

Finally, we examine:

- *How* can we learn from these practices, share knowledge, and leverage innovation to obtain similar results more broadly?

Teachers across the country are familiar with Professional Learning Communities (PLCs) where discussions about individual student success are common. PLCs underscore the importance of collective ownership, where we are all accountable for the learning and progress of all our students. The tool enables users to have meaningful conversations across different levels (district, school, department, teacher) using credible evidence depending on their access levels.

These questions and the ones that follow are intended to provoke and promote discussions on individual student progress over time. The focus is not on attaching rewards to teachers at this time but to expeditiously identify best practices already happening in the schools based on the growth and achievement trends of students in these teachers' classes. The primary goal of our work is to completely respect and support the hard work being done daily by the teachers and administrators in the trenches without promoting competition or creating differences between them (Norville, 2009). Just as students have a goal of reaching and exceeding the standards, we want our teachers to excel and have all their students exceed pre-established criterion on growth and achievement using our quadrant model (Figure 2).

Possible Modes of Inquiry

Before proceeding to the second phase, we must stress that the key to any school improvement effort aimed at changing instructional practice is the commitment to professional development (Balasubramanian, Frieler, & Asp, 2008). Teachers across the country meet periodically in Professional Learning Communities (PLCs) at their respective institutions to purposefully discuss how students learn (DuFour & Eaker, 1998). When student learning is the focus of a PLC, the data inquiry usually focuses on four questions:

Q1: What do we want our students to know and be able to do?

Q2: What evidence do we have that they have learned it?

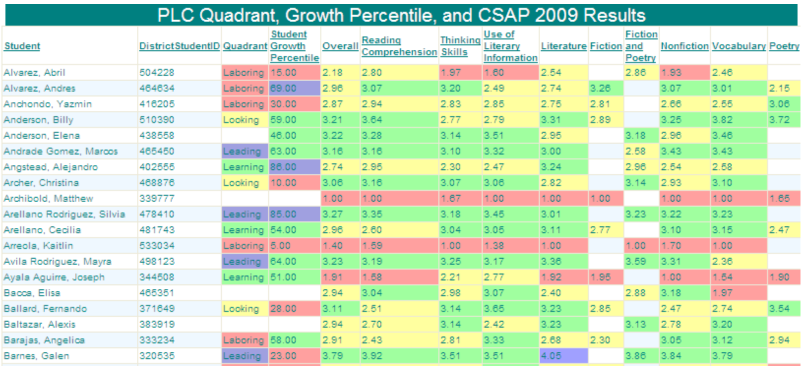
Q3: What do we do when students have not learned it (SW quadrant)?

Q4: What do we do with students who have already learned it (NE quadrant)?

The quadrant model (Figure 2) contributes two necessary and essential questions that were absent until now in PLC conversations. These are—Q5: How do we engage and promote learning among high-achieving students who are doing worse this year than last year (to move them from the NW quadrant to the NE quadrant), and Q6: How do we develop proficiency among low achieving students who are doing better this year than last year (to move them from the SE quadrant to the NE quadrant). With these two questions, teachers can identify the "lowest hanging fruit" for effective data inquiry and achieve improved results.

Additionally, from the visual representations of the quadrant plots, educators can drill-down into the strengths and needs of individual and groups of students using the "Strength Charts" (Figure 4) using the second tool in HarnessData®. These charts are useful to further conversations started with the PLC Quadrants. To help teachers and administrators see the improvements they are making in context, the strength charts have summary averages by teacher, school and district.

The column labeled “quadrant” in Figure 4 presents the value-added growth of the individual students. For example, the first student is in the “looking” quadrant. This means this high-achieving student has lost ground this year in the annual test when compared to their own performance last year. The second column titled “student growth percentile” (SGP) is the calculated growth metric for each student (and provided by the Colorado Department of Education). The SGP (13, for example) is computed by comparing individual student performance with their peers who started at the same scale score last year. The low student growth percentile (CDE, 2009a) shows this student’s normative growth. The third column “overall” displays the achievement level of the students relative to the state standard. It is criterion-referenced. For example, the 3.01 indicates the student is barely proficient. By examining these columns with students’ value-added, normative and criterion-referenced performance, educators can better understand assessment data. Additionally, by studying the performance of a student in the content standards and sub-content areas, the strengths and needs of each student can be diagnosed. Targeted individualized instruction and supports can be arranged. By clicking on each of these variables (quadrant, SGP, overall, content standards and sub-content areas) they can be further sorted to plan for differentiated instruction for groups of students excelling or struggling in the same variables.



PLC Quadrant, Growth Percentile, and CSAP 2009 Results

Student	DistrictStudentID	Quadrant	Student Growth Percentile	Overall	Reading Comprehension	Thinking Skills	Use of Literary Information	Literature	Fiction	Fiction and Poetry	Nonfiction	Vocabulary	Poetry
Alvarez, Abril	504228	Laboring	15.00	2.18	2.80	1.97	1.60	2.54		2.86	1.93	2.46	
Alvarez, Andres	464634	Laboring	69.00	2.96	3.07	3.20	2.49	2.74	3.26		3.07	3.01	2.15
Anchondo, Yazmin	416205	Laboring	30.00	2.87	2.94	2.83	2.85	2.75	2.81		2.66	2.55	3.06
Anderson, Billy	510390	Looking	59.00	3.21	3.64	2.77	2.79	3.31	2.89		3.25	3.82	3.72
Anderson, Elena	438558		46.00	3.22	3.28	3.14	3.51	2.95		3.18	2.96	3.46	
Andrade Gomez, Marcos	465450	Leading	63.00	3.16	3.16	3.10	3.32	3.00		2.58	3.43	3.43	
Angstead, Alejandro	402555	Learning	86.00	2.74	2.95	2.30	2.47	3.24		2.96	2.54	2.58	
Archer, Christina	468876	Looking	10.00	3.06	3.16	3.07	3.06	2.82		3.14	2.93	3.10	
Archibold, Matthew	339777			1.00	1.00	1.67	1.00	1.00	1.00		1.00	1.00	1.65
Arellano Rodriguez, Silvia	478410	Leading	85.00	3.27	3.35	3.18	3.45	3.01		3.23	3.22	3.23	
Arellano, Cecilia	481743	Learning	54.00	2.96	2.60	3.04	3.05	3.11	2.77		3.10	3.15	2.47
Arreola, Kaitlin	533034	Laboring	5.00	1.40	1.59	1.00	1.38	1.00		1.00	1.70	1.00	
Avila Rodriguez, Mayra	498123	Leading	64.00	3.23	3.19	3.25	3.17	3.36		3.59	3.31	2.36	
Ayala Aguirre, Joseph	344508	Learning	51.00	1.91	1.58	2.21	2.77	1.92	1.95		1.00	1.54	1.90
Bacca, Elisa	465351			2.94	3.04	2.98	3.07	2.40		2.88	3.18	1.97	
Ballard, Fernando	371649	Looking	28.00	3.11	2.51	3.14	3.65	3.23	2.85		2.47	2.74	3.54
Baltazar, Alexis	383919			2.94	2.70	3.14	2.42	3.23		3.13	2.78	3.20	
Barajas, Angelica	333234	Laboring	58.00	2.91	2.43	2.81	3.33	2.68	2.30		3.05	3.12	2.94
Barnes, Galen	320535	Leading	23.00	3.79	3.92	3.51	3.51	4.05		3.86	3.84	3.79	

Figure 4. Strength charts, displaying the mastery and needs of individual students

Note that the unit of analysis is the individual student in both the quadrant model (Figure 2) and the strength charts (Figure 4). Further, using these two diagnostic and continuous improvement tools from HarnessData®, teachers (as researchers) might use the following questions as a protocol to guide PLC conversations for continuous improvement:

- What instructional strategies seemed to work and did not work for me?
- Which groups of students (in the six subgroups) improved (or declined) in each performance level?
- What do we notice about the growth of these students (individually and collectively) in each performance level?
- Which groups of students pose additional questions?
- How does my performance relate to that of my colleagues who teach the same course within the school and across the district?
- What does student work look like?
- Do we have exemplars to share with our colleagues?

- Informed by this analysis, what is my plan of action for the next/current school year in terms of instructional interventions for students (individually and collectively) in the four quadrants?

Administrators (as evaluators) might ask:

- What evidence do we have to determine which programs and interventions are working, which ones are not, which ones need adjusting, and who needs support?
- How can we leverage resources to maximize the learning opportunities for all students?
- What performance indicators could we use to evaluate the effectiveness of processes and implementation across the district?
- Looking back at last year's school/district improvement plan and available data (on formative and summative school, district, and state tests), can we reflect on them—for celebrations, roadblocks, opportunities—and work ahead to address problem areas?
- What resources, including professional development and training, might be necessary to bring about sustainable learning district-wide?

The first phase of the project we discussed utilizes the Who→What→How process using annual standardized test data. It provides accessible information from test data and has strong technical support. However it tends to be narrow in its scope because the standardized tests we use currently are limited to reading, writing, mathematics and some science for students in grades three through ten. And the costs of tests continue to grow each year. In 2005 alone, these assessments generated $2.8 billion for the testing industry (Glovin & Evans, 2005). Regardless of cost and validity, these tests are still useful because they provide enough signals and information that can be gleaned from them when examined at the student and teacher levels. However, the results displayed are not representative of the comprehensive all-round educational programming provided by a school. They also do not measure the total worth of a teacher.

Teachers relate to their students daily at levels far deeper than what our current assessments can measure. They spend a considerable amount of time instructing and modeling critical learning and life skills for postsecondary and workforce readiness (CDE, 2009b), which are: critical thinking and problem-solving; finding and using information/information technology; creativity and innovation; global and cultural awareness; civic responsibility; work ethic; personal responsibility; communication; and collaboration. Regardless, we can observe both positive and negative trends across departments and teachers, using an incremental and standardized scale. On this given scale, we are going to look for success stories to share, learn from and use it to improve our professional practice. At the present time, our studies are exploratory and action-based in nature. We look forward at some point to sharing the results from more thorough investigations.

The second phase of the project takes us beyond the bare-bone student learning outcomes in reading, writing, math and some science. However, it is more labor intensive and culturally dependent on the values and intentions of the participants. We recommend that the overall model be implemented in two phases. Using the first phase, schools and districts can gain confidence in decision-making using test data. Once comfortable in the first phase, as they see the value and results (and perhaps the need for a more complete picture of outcomes) these institutions can then engage in the second phase. This phase requires more inquiry and co-development where the local participants really need to make meaning by using it frequently to monitor student learning over time. This local adoption is designed to link and bridge diagnostic and formative assessments used at the school with state annual tests. The second phase presented below is fully articulated. The findings from three case studies that we are currently in the midst of using Phase I will be the subject of another paper.

Phase 2: Evidence of Learning and Progress Monitoring

To promote credible evidence-based approaches using this feedback and decision-making tool, the second phase of the project starts with the "where," then into the "when," and lastly the "why." These three prompts along with the earlier three prompts present a comprehensive view of how a systems approach can make a difference to educational outcomes in both the short-term and long-term. This phase of the project is a data-rich linking of local diagnostic and formative school and district assessments to the results on the standardized tests. With our ongoing "Evidence of Learning" tool (Figure 5 and third HarnessData® tool), substantive conversations on individual students and their overall progress can be supported within departments (vertical teams across grade levels) and across departments (horizontal teams at specific grade levels). These cross departmental collaborations are not limited to the narrowness of the existing standardized tests but could be more broad-based for providing systemic support to all students.

The Where→When→Why Process

As researchers, we begin by asking:

- *Where* do struggling students congregate within a school (based on evidence observed from student enrollment records)?

With help from the administrators, we then ask:

- *When* can the core subject teachers collaborate with the encore (or elective subject) teachers?
- Which instructional strategies and interventions do they frequently utilize?

Finally, we examine:

- *Why* not problem solve as a system of three or four departments (primarily teams of teachers in these departments who instruct students for at least nine weeks) who might collectively impact individual student performance in the state reading, writing, mathematics and science tests?

The ongoing "Evidence of Learning" tool displays the results of students from each of the four quadrants (Figure 5) in the school/district diagnostic and formative assessments over time. The unit of analysis is the collective means of the four groups of students in the PLC quadrants (leading, learning, looking and laboring). In the example shown one can observe the green and orange lines crossing each other. The display illustrates the "catching up" effect where underperforming "learning" students (green line starting at 38%) are catching up when their high-performing "looking" (orange line starting at 50%) counterparts. Can we create a learning culture within the organization by leveraging such noticeable changes in how students progress over time? The high performing "looking" students display an almost flat growth rate. If not attended to, some of these "looking" students will begin to slip more and even drop to become more underperforming students without immediate intervention. From September to October (see Figure 5), we can see that the "looking" students are losing ground even in the interim school assessments. They seem to be on track to do worse this year than last year in the annual tests if they are not engaged with appropriate instructional interventions.

For instance, Sally, John and Jorge might be low performing "laboring" students in their core subjects (contributing to the collective mean of the red line starting at 27% in Figure 5). They however happen to be friends all taking band. How can we use their musical interests to leverage and propel them to become successful on the basic measures of reading, writing, math and science tests? If we can work from this strength that they are there in school together and are still connected to school, can we have these three students say, demonstrate their competencies and interests on a performance-based assessment in music to problem solve at the individual student level? This approach works against the thinking that we are all the same and must conform to going through a standardized factory model of education. Instead it exploits the differences in the individuals and recognizes that we are not trying to get everybody to become a "computer jock," or "nerd" or "musician," say. We are going to let these differences accentuate and work from some strengths while maintaining the achievement and growth levels of all the students in the core subjects. This way, engagement and achievement through engagement becomes the goal of schools.

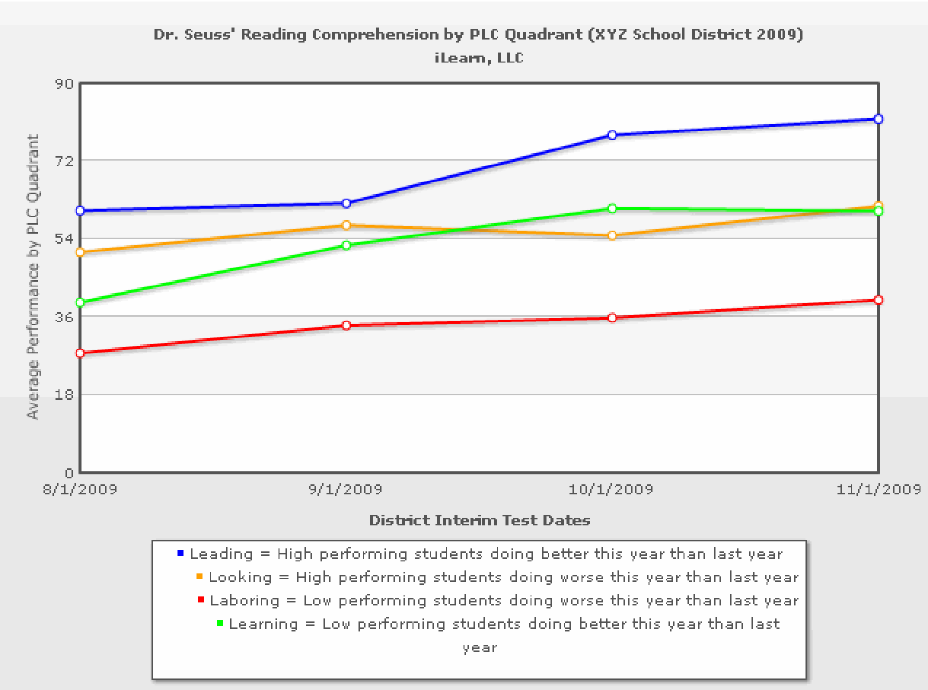


Figure 5. Monitoring the progress of students across the different quadrants in the school/district assessments

Schools might find, for example, that the art or technology classroom may be the haven for some of the non-fitting students who are hard-pressed and unmotivated to "do" and finish school in the traditional sense (with a focus on only core classes). The elective or encore classes might be the place where these individuals' passions, interests and the connections with each other can be established. This community and belonging in itself can help these students achieve and be motivated in all the classes (both core and encore). It is part of keeping these students engaged, connected and on task with the assignments in the core classes. The ongoing "Evidence of Learning" tool allows teachers in these "encore" subject areas along with the counselors to still be involved in school-wide conversations of underperforming and high performing students in the four quadrants.

Possibilities and Pitfalls

Several possibilities open up when effective teaching can be recognized. The PLC conversations become more concrete and contextualized with credible evidence from a tool like HarnessData®. The data on student growth and achievement can be better understood when examined: relative to the student's prior performance (value-added), relative to their peers (normative), and relative to the standards (criterion-referenced). Schools can examine instructional alignment (how and when students are taught); identify what strategies and working, with whom, and what needs adjustment; validate effective instructional strategies; recognize which teachers consistently demonstrate high effectiveness across the different subgroups, content standards, sub-content areas, growth and achievement variables. Schools can link student performance on annual state tests to district/school diagnostic and formative assessments across all subjects. Districts can examine curricular alignment (what students are taught); look where they can find good instruction; identify which populations of students are and are not being served well; discover which programs and strategies are being effective and more.

While promising, stakeholders should interpret the results of all assessment data with some care. Still more work needs to be done as we co-develop training materials with educators and integrate this approach into school-improvement efforts during our pilot case studies. The interpretations might be skewed by measurement errors associated with testing. These errors are larger on the sub-content areas compared with the content standards because of fewer items assessed on each sub-content area. In fact, the sub-content areas are designed to provide more diagnostic than summative information; Other factors that might impact these interpretations are regression toward the mean, test-ceiling effects for some high achieving students, and improvements in test taking skills associated with repeated, within-year assessments. Schools and districts should be wary of other common pitfalls. Schools should not think that all points of growth on this value-added scale are identical; or focus only on the "bubble students;" or teach only to the item formats on large-scale assessments; or teach to just the content areas on these assessments. Districts should not assume that there are no errors in measurement; or focus the resources on only the below proficient students; or rely just on state tests alone to measure instructional and program effectiveness.

Conclusion

Improving the performance of students with diverse needs and abilities amidst the data deluge in K12 education can be a challenge. This paper introduced a web-based application to visually represent test performance of groups of students based on two factors: criterion-referenced achievement and value-added growth. The application can be used to provide constructive and timely feedback to teachers and administrators that is intuitive and does not require them to be familiar with advanced statistical analyses. It can be used to identify gaps in learning and inform classroom instruction. The particular analytic method presented here allows for identifying different trends among subgroups of students using the quadrant model and strength charts. In contrast to the oversimplified approach of tracking a single group mean, the tools presented in this article analyzes the data in a way that facilitates a more tailored approach to interpreting assessment data. This can support decisions about educational strategies and interventions that are appropriate to subgroups of students based on their academic achievement. In Figures 2, 4 and 5 we introduced three powerful HarnessData® diagnostic and continuous improvement tools and how they can be used to track progress toward learning goals and targets on achievement and growth.

Additionally, the quadrant model contributes two necessary and essential questions for effective data inquiry using PLCs: How do we engage and promote learning among high-achieving students who are performing worse this year than last year? And, how do we develop proficiency among low achieving students who are doing better this year than last year? By building on the existing framework of PLCs and making the connections explicit, educators can see its utility.

The tools in HarnessData® are designed to provide educators with an accessible and meaningful form of data analysis that can be used by teachers within PLC discussions. The tools may also stimulate the thinking of administrators (and evaluators) working in educational settings about how they might easily share assessment data with stakeholders. Our hope is that tools like HarnessData® will lead to better discussions, better strategies, better interventions, better instruction, and consequently better learning, which will in turn help sustain system-wide continuous improvement practices.

References

Babu, S., & Mendro, R. (2003, April). *Teacher accountability: HLM-based teacher effectiveness indices in the investigation of teacher effects on student achievement in a state assessment program*. Paper presented at the Annual Meeting of AERA, Chicago, IL.

Balasubramanian, N, Frieler, J., & Asp. E. (2008) Designing learning. *Principal Leadership*, *9*(2), 34-39.

Balasubramanian, N, & Bankes, P. (2009, November). *Using longitudinal assessment data: Feedback tool for teachers and students*. Demonstration session presented at the Annual Meeting of the American Evaluation Association, Orlando, FL.

Balasubramanian, N., & Muth, R. (2010. April). *Instructional and program effectiveness: From data to action*. Paper presented at the AERA Annual Meeting, Denver, CO.

Barton, P. E. (2009). *National educational standards: Getting beneath the surface*. Princeton, NJ: Educational Testing Service.

Black, E. (2001). *IBM and the holocaust: The strategic alliance between Nazi Germany and American's most powerful corporation*. New York: Crown Publishers.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice, 28*(4), 3-14.

Bush Foundation. (2009, December 3). Press conference (Video recording). *Teacher Effectiveness Initiative of the Bush Foundation and its 14 partner Universities*. Bush Foundation, Saint Paul, MN.

Colorado Department of Education (2008). *Performance Level Scale Ranges for CSAP Assessments*. Retrieved from

http://www.cde.state.co.us/cdeassess/documents/csap/manuals/2008/20080718_scalescoreranges.pdf

Colorado Department of Education. (2009a). The Colorado growth model: Frequently asked questions.

Retrieved from http://www.schoolview.org/documents/CGM_FAQ.pdf

Colorado Department of Education and Department of Higher Education. (2009b). *Postsecondary and Workforce Readiness Definition*. Retrieved from http://www.cde.state.co.us/cdegen/downloads/PWRdescription.pdf

CTB McGraw-Hill (2009). CSAP technical report 2009: Submitted to the Colorado Department of Education. Monterey, CA: CTB McGraw-Hill. Retrieved from http://www.cde.state.co.us/cdeassess/publications.html

DuFour, R., & Eaker, R. (1998). *Professional learning communities at work: Best practices for enhancing student achievement*. Bloomington, IN: National Education Service.

Glovin, D., & Evans, D. (2006). How test companies fail your kids. *Bloomberg Markets*. 126-138.

Kipling, J. R. (1912). *Just so stories for little children*. New York: Doubleday.

Lewis, D. M., Green, D. R., & Mitzel, H. C. (1998, April). *The bookmark standard setting procedure: Methodology and recent implications*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Mathison, S. (2009). Seeing is believing: The credibility of image-based research and evaluation. In S. I. Donaldson, C. A. Christie, & M M. Mark (Eds.). *What counts as credible evidence in* *applied research and evaluation practice* (pp. 181-196). Thousand Oaks, CA: Sage Publications.

Millman, J. (Ed.). (1997). *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* Thousand Oaks, CA: Corwin Press.

Norville, D. (2009). *The power of respect: Benefit from the most forgotten element of success*. Nashville, TN: Thomas Nelson.

Perie, M., Weiss., J., Kurtz, T., & Dunn, J. (2009, June). *Comparing the use of a growth model and an index in a state accountability system*. Session 041 presented at the meeting of the National Conference of Student Assessment, Los Angeles, CA.

Popham, W. J. (1997). The moth and the flame: Student learning as a criterion of instructional competence. In J. Millman (Ed.). *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 264-274). Thousand Oaks, CA: Corwin Press.

Reeves, D. B. (2006). *The learning leader: How to focus school improvement for better results*. Alexandria, VA: Association for Supervision and Curriculum Development.

Author Notes

Nathan Balasubramanian, Ph.D., is President of iLearn, LLC. Brent G. Wilson, Ph.D., is professor of Information and Learning Technologies at the University of Colorado Denver. Correspondence should be addressed to Dr. Nathan Balasubramanian, iLearn, LLC. 651 Homestead Street, Lafayette, CO 80026. Phone: (720) 936-5999 E-mail: nathan@iLearnLLC.com.

The HarnessData® tool can be accessed by anyone interested at https://HarnessData.org with a username and password. Please e-mail the first author to request a username and password.