

DRAFT REPORT-- Do not cite or distribute without permission of authors

Building With Benchmarks:

The Role of the District in Philadelphia's Benchmark Assessment System

Katrina Bulkley
Montclair State University

Jolley Bruce Christman
Research for Action

Margaret E. Goertz
University of Pennsylvania

Nancy R. Lawrence
University of Pennsylvania

Paper presented at Session 60.041: Routine Checkups: What's the Prognosis for Improving
Schools and Student Learning with Interim Assessments?

Annual Meeting of the American Educational Research Association
New York, New York

March 27, 2008

Research for this paper was supported by a National Science Foundation Grant REC-0529485 to the Consortium for Policy Research at the University of Pennsylvania and by grants from the Spencer Foundation and the William Penn Foundation to Research for Action. The opinions expressed in this research are those of the authors and do not necessarily reflect the views of the Consortium for Policy Research in Education, the National Science Foundation, the Spencer Foundation, the William Penn Foundation, or the institutional partners of CPRE.

Introduction

In recent years, as the push for increased achievement has intensified, districts are increasingly turning to “interim assessments” to track students’ progress at regular intervals throughout the year. While some districts are writing their own benchmark assessments, others are turning to the significant for-profit industry that is springing up to sell districts these assessments and the technology needed to administer and score them and analyze results (Burch, 2005). More and more districts are investing in expensive data management systems designed to help teachers, principals, and district leaders make sense of student data, identify areas of strength and weakness, identify instructional strategies for targeted students, and much more (Olson, 2005).

These new assessments, often called “benchmark assessments,” are given multiple times a year in various subjects. According to a 2005 article in *Education Week*, 7 of 10 superintendents surveyed give district-wide tests, and another 10% said that they planned to give such tests in the following year (Olson, 2005). Most of these tests are designed to predict students’ performance on end-of-the-year state exams that serve as an important measure in determining whether a school makes its Annual Yearly Progress (AYP) target. However, there are some districts that have implemented interim assessments that are more closely aligned to a district-wide curriculum or district standards and have been designed for the purpose of providing educators with formative information about students’ mastery of the curriculum (Olson, 2005).

However, we know little about how school-based educators use results from benchmarks to assess student understanding and modify academic programs, or about the conditions that support their ability to use such information. Indeed, there is considerable debate in both the policy and education communities about whether benchmark assessments *can* be used formatively; that is, to inform and direct teachers’ instruction on a regular basis. We also know little about the kinds of policies and supports that might facilitate more formative use of these kinds of assessments.

Philadelphia provides an excellent case study for exploring these issues. The district has been an active user of districtwide benchmark assessments since September of 2004, with a subset of low-performing schools using them prior to this date. The benchmark assessments were part of a larger reform initiative in the district, and are designed to be aligned with district standards and curriculum. While other papers in this symposium focus on how teachers use benchmark assessments, this paper examines the district role in the design and implementation of benchmarks in Philadelphia. Specifically, four research questions guided the analysis:

1. What is the accountability, organizational, and instructional context for benchmark assessments in Philadelphia?
2. What are the district leaders’ expectations concerning the use of the benchmark assessments?
3. What supports do the district and the education management providers provide for use of the benchmark assessments and instructional improvement and how are they used in schools?
4. What are the challenges to meeting district leaders’ expectations for use of the benchmark assessments to improve instruction?

We begin the paper with a discussion of where benchmark assessments are situated in the continuum of summative-to-formative assessments. The second section presents the

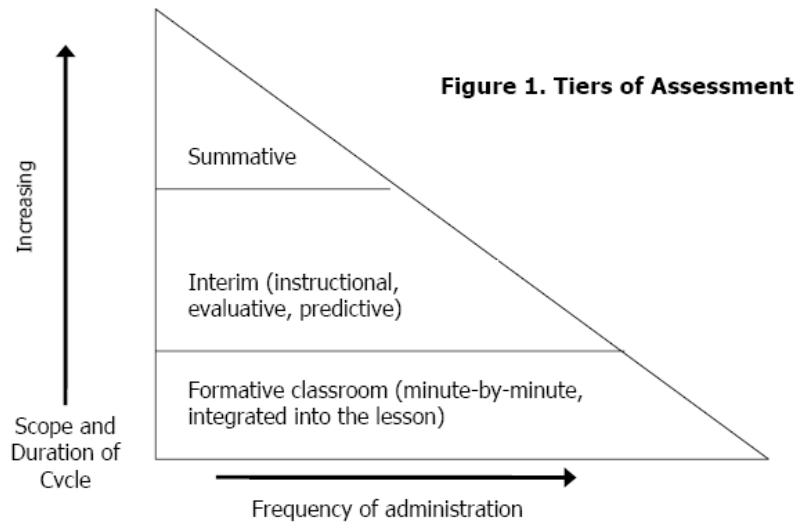
methodology for our study. The third section documents the policy context for the development of the benchmark assessments, while the fourth section describes the assessments themselves. The fifth section discusses the district's expectations for the use of the assessments; the sixth section outlines district supports for their use. The paper concludes with challenges to meeting the district's expectations for the use of the benchmarks.

What Are Benchmark Assessments?

At the April 2007 American Educational Research Association's annual meeting in Chicago, 177 sessions were devoted to the topic of assessment. Included in this broad category were discussions about "formative assessments," "benchmark assessments," and "interim assessments." Reflected in these sessions was an on-going debate about what benchmark assessments are and whether they are or can be formative in nature.

Perie et al. (2007) have categorized the three kinds of assessments currently in use—summative, formative, and interim—by their intended purposes, audiences, and the frequency of their administration. (See Figure 1.)

- Summative assessments are given at the end of a semester or year to measure students' performance against district or state content standards. These standardized assessments are often part of an accountability system and are not designed to provide teachers with timely information about their current students' learning.
- Formative assessments occur in the natural course of teaching and learning. They are built into classroom instructional activities and provide teachers and students with ongoing, daily information about what students are learning and how teachers might improve instruction so that learning gaps and misunderstandings can be remedied. These assessments do not provide information that can be aggregated.
- Interim assessments "fall between formative and summative assessment" and provide standardized data that can be aggregated. Interim assessments vary in their purpose. They may predict student performance on a summative, accountability assessment; they may provide information for an evaluation of a program; or, they may diagnose student strengths and weaknesses.



Source: Perie et al. (2007)

Black and Wiliam’s meta-analysis (1998) established the powerful positive effect of “formative” classroom assessments that offer teachers “minute by minute” checks on student learning because they are embedded in instructional activities. But their view of formative assessment focuses on purpose as well as frequency, and they define “formative” as “encompassing *all* those activities undertaken by teachers, and or/by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged” (emphasis added). Perie et al. (2007), however, argue that the current “interim” assessments are not formative in the true sense of the term as defined by Black and Wiliam and should not be viewed as yet having the same evidence base to support their use, despite claims by vendors and districts themselves.

Yet, Perie et al. (2007) also recognize the formative properties of interim assessments, and the purposes that they serve beyond the individual classroom:

Interim assessments (1) evaluate students’ knowledge and skills relative to a specific set of academic goals, typically within a limited time frame, and (2) are designed to inform decisions at both the classroom and beyond the classroom level, such as the school or district level. Thus, they may be given at the classroom level to provide information for the teacher, but unlike true formative assessments, the results of interim assessments can be meaningfully aggregated and reported at a at a broader level (p. 3).

And Halverson, Pritchett, and Watson (2007) warn that the summative/formative dichotomy is overly simplistic.

The distinction between summative and formative often lies in the perception of the communicators, not in the information itself. Thus, information generated, for example, through shared assessments or peer observation can be interpreted and

used as evidence to summatively judge and discipline teachers, just as standardized text scores can be used to formatively reshape instructional practices (p. 5).

As we will see in a later section of this paper, benchmarks in the School District of Philadelphia have been used in both summative and formative ways.

Methods

The findings reported here are drawn from interviews of district leaders and school principals in Philadelphia that were conducted by researchers from the Consortium for Policy Research in Education (CPRE) and Research for Action (RFA) during the 2005-06 and 2006-07 school years.

CPRE study

In the winter and spring of 2006, CPRE interviewed five district leaders, including regional superintendents and curriculum, assessment, and technology leaders from Central Office. The goals of these district interviews were to gather information on the benchmark assessment system, the district's expectations for the benchmarks, the mathematics curriculum (*Everyday Mathematics*), data analysis and use, and professional development. In the spring of 2006, CPRE interviewed the principals of six elementary schools in the study. A central purpose of these interviews was to identify district- and school-level expectations for benchmark assessment use and potential supports offered to teachers. CPRE's school sample included six schools that had made AYP and had scored at or above average on the state's mathematics assessment. These schools mirrored the demographics of the district. All schools were Title I schools; four schools were 90-99% African-American, and the other two schools were approximately 99% Latino. The six schools were located in three different regions across the district, each headed by a Regional Superintendent and supported by a regional staff.

In addition, CPRE researchers attended professional development sessions, technology training sessions, and principal meetings (SchoolStat meetings), during which several types of "performance indicators," including benchmark results, were discussed. CPRE also collected the actual mathematics benchmark assessments and other relevant documents, such as district curriculum and instruction guides. All interviews were transcribed and coded in Atlas.ti to identify major themes and similarities that cut across the district and schools.

RFA Study

Included in the analysis for this paper were interviews by RFA researchers of five administrators from the Office of Accountability, Assessment, and Intervention, Office of Curriculum, and Office of Professional Development. These interviews occurred during the 2005-06 school year, were semistructured, lasted approximately one hour. The topics covered included the core curriculum, student performance assessments, generally, as well as in-depth probing about benchmark assessments, professional development for school leaders on using data, and perceptions of if and how the Education Management Organizations operating in the district were making use of the district's core curriculum and benchmark system.

Additionally, during 2005-06 and 2006-07, we conducted semistructured interviews with the principals in 10 schools that were part of our sample. One school underwent a leadership transition during the course of the study. Five of the 11 principals had been principals for less

than three years. Each principal was interviewed at least twice, and the majority three times. Each interview lasted approximately 90 minutes and covered such topics as the kinds of student performance data their schools routinely examined, their own as well as other school leaders' role in the examination of data, and the supports their schools received from their provider organizations as well as from the district. All interviews were transcribed and coded in Atlas.ti to identify major themes and similarities that cut across the district and schools.

RFA's school sample included 10 elementary schools that had been identified as "low performing" and eligible for intervention under state takeover. Seven of the schools were under management by outside providers; two schools were part of the district's homegrown intervention under the Office of Restructured Schools; one school was a "sweet sixteen" school—a low-performing school that was showing improvement and therefore received additional resources but was not under a special management arrangement. Each of the 10 schools was nominated either by its outside management provider or by the district as a school that was a good example of its approach to the use of data to inform decision making. All of our schools served a high percentage of students living in poverty. (The median percentage was 86.3%, considerably higher than the district average.) As mentioned, seven of the schools served student populations that were more than 90% African American. The three other schools had substantial Latino populations.

The Context for Benchmark Assessments in Philadelphia

Since a state takeover in 2001, the School District of Philadelphia has served as a laboratory for fundamental changes in school governance and management, most notably a complex privatization scheme that includes market solutions such as a "diverse provider" model of school management, expansion of charter schools, and extensive outsourcing of additional district functions.¹ At the same time, they instituted strong centralizing measures including a district-wide core curriculum, mandated after-school programs, and conversion of middle schools to K-8 schools. They also made the district a national frontrunner in welcoming the spirit and accountability mechanisms of the 2001 federal No Child Left Behind Act (NCLB). The district has combined what Wong and Shen describe as the leading—and many would say, contradictory—alternatives for reform strategies; namely, market-based solutions along with a strong centralized authority model (Wong & Shen, 2003).

Curriculum, Assessment and Accountability

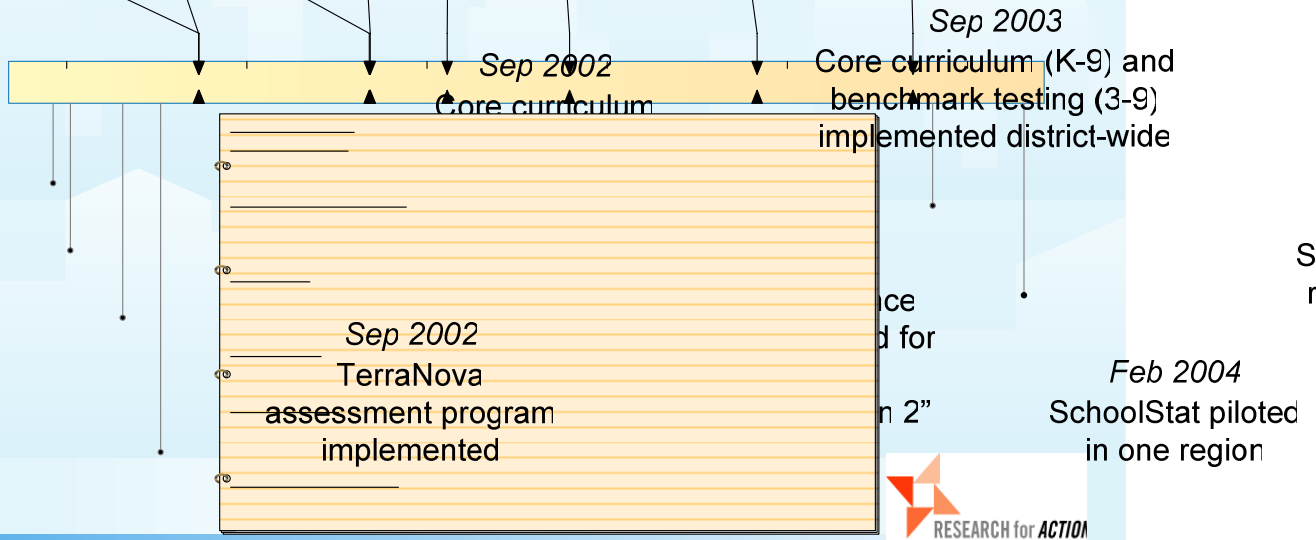
Districtwide assessment has been an integral part of Philadelphia's education reforms for over 25 years. Over this time, test results have been used for both accountability and instructional purposes. The centerpiece of Superintendent Constance Clayton's 12-year administration (1980-92) was the K-12 Standardized Curriculum, which was accompanied by a pacing guide that laid out a week-by-week schedule for instruction and a criterion-referenced test aligned with the district curriculum and administered annually. When David Hornbeck became superintendent in 1994, he brought standards-based reform to Philadelphia. The district abandoned the Standardized Curriculum and the citywide tests as emphasis shifted from teachers covering a prescribed curriculum to all students meeting rigorous performance standards. Although a curriculum did not initially accompany new locally developed standards, in 1998, in response to requests from school-based educators, district leaders issued the Curriculum Frameworks, which offered teachers some curricular guidance. In Philadelphia's first move toward accountability based on student achievement, the district adopted the SAT9. The test

became an important part of a Performance Responsibility Index as principals' performance reviews and salaries were tied to their schools meeting district-established targets.

In 1998, the governor and legislature passed the first of two bills that enabled a state takeover of the district (Maranto, 2005; Boyd & Christman, 2003). In December of 2001 the district and state compromised on a "friendly takeover," that included a new School Reform Commission (SRC). Three of the SRC's members are appointed by the governor, and the remaining two are appointed by the mayor. Six months later, the SRC hired former Chicago Public Schools CEO Paul Vallas to head the Philadelphia district.

As shown in Figure 1, one of Vallas' first initiatives was to institute a districtwide core curriculum in four academic subjects for grades K-8. In 2003-04, the district added a requirement that all elementary students have 120 minutes of literacy and 90 minutes of math per day, based on the core curriculum (Travers, 2003). Increased testing, including the six-week formative benchmark tests, also accompanied the new core curriculum (Useem, 2005). Alongside the creation of the K-8 core curriculum (which was mandatory for regular district schools but voluntary for those managed by providers) was a move to eliminate middle schools in favor of K-8 schools. Vallas and the SRC also initiated a school quality review process, beginning in 2002-03; this process included "regular" district schools, those in the diverse provider model, and the city's many charter schools. Christman and her colleagues (Christman, Gold, & Herold, 2005) argue that, "Like the core curriculum, the school review process was a way for the district to exert its influence on providers and their schools" (p.12)

School District of Philadelphia Timeline



In addition, the district introduced two new performance and data management systems systemwide—SchoolNet and SchoolStat. In 2003, the district contracted with SchoolNet Instructional Management Solutions (SchoolNet) to organize and disseminate individual and aggregate benchmark assessment data and to make assessment data immediately accessible to teachers and principals. As discussed later in the section on District Supports, SchoolNet not only is a data management system, but provides a set of analytical and instructional tools aligned with the core curriculum to teachers, principals, and families.

SchoolStat was launched by the District and the University of Pennsylvania's Fels Institute of Government. During the 2003-04 school year, SchoolStat was rolled out to 15 elementary schools. By late spring of 2005, SchoolStat was expanded districtwide. SchoolStat compiles and compares school-level data on student performance and behavior and student and teacher attendance. This tool was used at regular meetings of Regional Superintendents and their principals and at meetings of the Regional Superintendents as a group with the District's

Jan 2002
NCLB signed
Into law

Apr 2002
Diverse providers
chosen by
School Reform
Commission

Jul 2002
Paul Vallas
appointed CEO

2003
TerraNova CTBS: Norm-referenced standard
Core Curriculum: A uniform curriculum for grades 3-9 (implemented in Sept. 2003); for grades 10-12 the curriculum (implemented in Sept. 2004); includes materials
School Assistance Team: Assigned to schools
of school-based, regional, and central office
process to help schools review their existing
habits around data use for continuous improvement
SchoolStat: A performance management system
1) data on student performance, attendance, and
meetings intended to help school leaders access
The SchoolStat contract was cancelled in summer 2005
Benchmarks: formative assessments administered
(administered less frequently in high schools)
implemented in grades 3-9 in September 2003
SchoolNet/IMS: Web-based instructional management system
performance data, curricular materials, PD materials
users include school staff, parents, and students
semester, with all schools equipped by March 2003
School Growth Teacher: Assigned to schools to
fully-released teachers who help improve academic

2004

Chief Academic Officer, to discuss the status of, and ways to improve, climate and achievement at their schools.

The Diverse Provider Model

While not initiated by Vallas, the creation of a diverse provider model (DPM) was probably the most visible change to result from the takeover in the landscape of Philadelphia public education (Bulkley, Mundell, & Riffer, 2004). Building on work by Paul Hill and his colleagues (Hill, Pierce, & Guthrie, 1997; Hill, Campbell, & Harvey, 2000) and the Edison report recommendations, the DPM was a response to a push from the state to create a more “market-based approach to the challenges facing Philadelphia public schools” (Bulkley et al., 2004, p.1).

In total, seven different organizations (three for-profit educational management organizations (EMOs), two locally based nonprofits, and two universities) were hired to provide some level of management services in 46 of the district’s 264 schools (Bulkley et al., 2004). To support these and other changes in Philadelphia, in July of 2002, the state provided \$55 million in additional funding to the district, \$37.5 million of which was devoted to the adoption of the diverse provider model. The SRC also created a separate Office of Restructured Schools (ORS) as its own internal “provider” to oversee 21 additional low-performing schools, granted additional funding to 16 low-performing schools that were making progress, and converted three additional schools to charter schools (Useem, 2005). Thus, the creation of the DPM initially affected 86 schools (Research for Action, 2005).

In theory, a diverse provider model involves shifting decision-making away from districts and towards outside managers. However, the schools and providers involved in the diverse provider model have felt the pull of the Vallas-initiated centralizing district reforms; as one provider’s representative commented, “Vallas is changing all the rules. We feel like we are being sucked into the tidal wave of centralized control” (quoted in Bulkley et al., 2004, p. 6). For example, many of the providers (with the exception of Edison) are using parts or all of the district’s core curriculum and the benchmark tests aligned with it.

As the date for a decision on the renewal of the EMOs’ five-year contracts approached in the spring of 2007, several studies on student outcomes were released. One of these studies was a longitudinal value-added analysis of individual student achievement conducted by RAND and Research for Action.ⁱⁱ It found that student performance gains in provider schools did not outpace student gains in other schools in the district, despite considerably greater resources. Further, the study found that the schools that were managed by the Office of Restructured Schools showed increased student gains in mathematics that outpaced the gains of the other providers and those of the rest of the district. Findings from the other two studies suggested that the diverse provider model might not have figured prominently in the district’s improvement and that the district itself had designed an intervention that showed promising results.

The School Reform Commission was divided about whether to extend the EMO contracts. In the end, despite vocal opposition from some parent and community groups, a three-member majority of the SRC voted to extend all of the contracts, albeit at a reduced per pupil expenditure and for only one year. Contracts will be up for renewal again in June 2008.

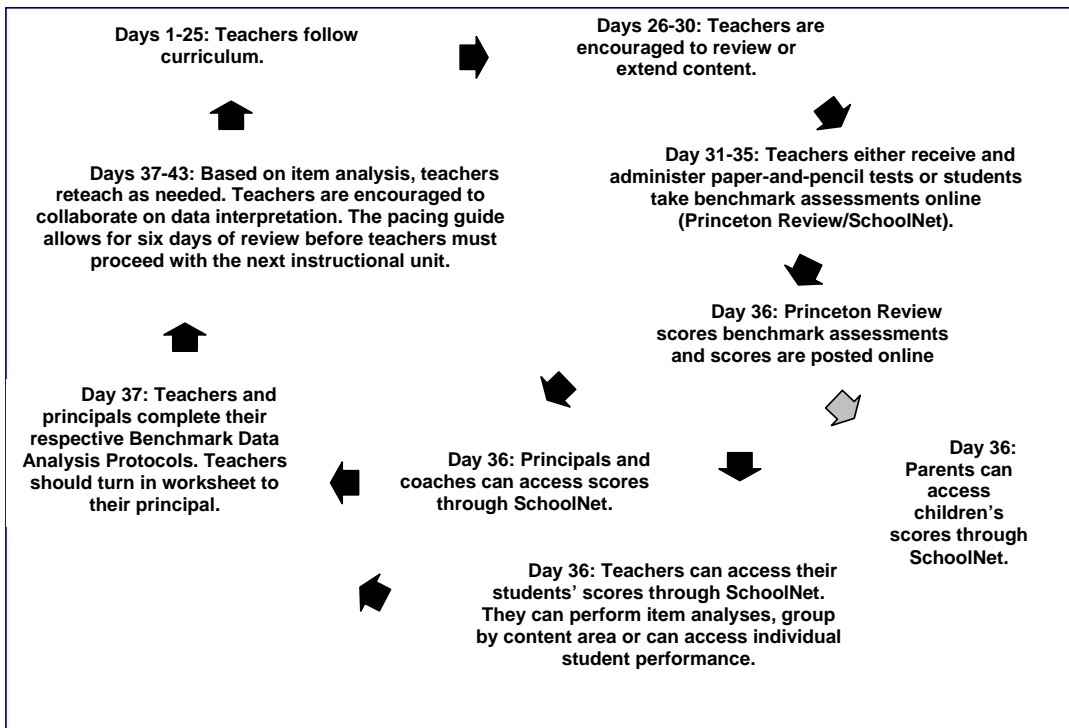
Current Context

Paul Vallas left the district for New Orleans in the summer of 2007. Two months after Vallas's departure, the chair of the School Reform Commission, James Nevels, a Republican appointee, also resigned. Sandra Dungee-Glenn, a mayoral appointee to the SRC and Democrat, assumed leadership. It was expected that SchoolStat would continue through August 2008, but when one of its most enthusiastic leaders left the district in 2007, the contract was not renewed. While some regional superintendents have tried to maintain the quarterly data reviews, the data are not as readily available. On February 19, 2008, the SRC appointed Arlene Ackerman, the former superintendent of the San Francisco and Washington DC school systems, the new CEO of the School District of Philadelphia.

The Philadelphia Benchmark Tests

Benchmarks assessments were implemented districtwide in Philadelphia in October 2004. Prior to that, they had been used in the set of schools managed by the district's Office of Restructured Schools (ORS) as part of the diverse provider model. The School District of Philadelphia (SDP) uses benchmark assessments in grades K-8 to give teachers immediate feedback relative to their students' mastery of the core curriculum topics taught in a series of cycles. As shown in Figure 2, each cycle of instruction and assessment consists of six weeks: five weeks of instruction at the end of which the benchmark assessments are administered and a sixth week of review and/or extended development of topics.

Figure 2: The cycle of instruction and assessment in Philadelphia



A review of accounts of benchmark use (largely in Education Week) led us to the conclusion that, in other districts, such tests are given between three times a year and monthly. Aside from Philadelphia, we did not identify any other districts where time was explicitly set aside for addressing weaknesses found through analyzing benchmark data.

In Philadelphia, each benchmark assessment is designed to test only those concepts and objectives taught since the last benchmark. All of the items in the benchmark assessments are multiple choice and come directly from the concepts and skills in the District's pacing guide, the Planning and Scheduling Timeline. The district administers benchmarks in Reading and Mathematics to students in grades 3-8. The Reading Benchmark contains 20 items; the Mathematics Benchmark also contains 20 items, except grade 8 has 25 items. Students in grades 3, 7, and 8 are given a Science Benchmark, containing 15 items.

The benchmark assessments are cocreated by the SDP curriculum and assessment administrators and Princeton Review in the weeks prior to their administration. According to central office administrators, this process requires several iterations for each of the five assessments given throughout the school year. The assessments are designed to be aligned to Pennsylvania's assessment anchors (and, therefore, to the content of the state test) and to the district's core curriculum and state standards. According to one district insider:

We wrote the standardized curriculum across the district and that was the first big step, and you know it's completely aligned to state standards and we broke it down into content descriptors and then we made sure it was at the proficient level... We divided the whole curriculum for the 3-8—the high schools quarterly, it's a little bit different—into six-week chunks so you teach for five weeks, you test exactly what was taught, you see whether or not the students learned what was taught, and then you take a week to reteach or enrich based on the data and that's really the premise.

District Expectations for Benchmark Assessment Use

The district's Office of Curriculum has identified multiple purposes for the benchmark assessments (School District of Philadelphia, 2007):

- To provide PSSA practice for students by simulating rigor, types of questions and building test-taking stamina;
- To provide teachers, administrators, students and parents with a quick snapshot of student progress;
- To determine if what is taught is what is learned;
- To help teachers reflect on instructional practices; and
- To provide data to assist in instructional decision-making.

While the district formally identified a number of purposes for the benchmarks, analysis of interviews with central office staff suggests two related purposes. First, the benchmarks would

provide formative feedback to teachers about their students' success in mastering concepts and skills covered in the core curriculum during the five-week instructional period. Second, the six-week cycle of teaching and assessment would, as one district leader noted, "create some kind of a pacing, and sequence, program." Over time, the latter purpose became less of a focus, as the enactment of the core curriculum following the district's expectations around pacing became institutionalized.

While central office leaders emphasized the use of benchmark data by individual teachers to understand what their students were learning, they also voiced expectations for other uses. Specifically, there were expectations for how a cycle of data analysis and improvement would "play out" on several different levels: the individual teacher, the instructional community (especially, grade groups), and the school. The nested nature of these cycles is shown in Figure 2. It was not always clear how information and ideas would flow between the different levels.

Expectations for Benchmark Use by Individual Teachers

At their core, the benchmarks were, in the words of several district leaders, "teaching tools"—in other words, tools that would support teachers' instruction by providing them with timely information about what their students were learning. The straight-forward theory of action underlying the benchmark instruction and assessment cycle was that a teacher would teach subject matter from the core curriculum for five weeks. During the fifth week, students would take the benchmark test. Teachers would analyze results from the benchmark test and, based on this information, would design a sixth week of instruction that would provide remediation for students in areas of weakness and enrichment in areas of strength. Consequently, students would make more progress towards mastering the concepts taught. After the sixth week, the cycle would begin again with new subject matter. One administrator described why this would be helpful to teachers:

We started with benchmarks 'cause that's the only formative piece we have, you know that became the one big thing that teachers had where they could change directions if they needed to make midcourse corrections, whereas before you wait every year for PSSA, you use it the best you can and it's not even the students sitting in front of you anymore, so right now with SchoolNet you can get data of the students sitting in front of you today, not the kids who went on to a different school from last year. So it's, you know, dynamic; it's not static anymore as far as teachers being able to see the kids in the class.

When asked, "What would it [use of benchmark data] look like hands-on for teachers in the ideal?", a district leader responded:

Well, we've asked them, "How might you regroup students on certain skills and knowledge?" Like, you say, you know what, there's a whole group of kids who missed this, but there's another group that really got it. And most of our standard statements, there's more than one question, so we can tell they either got them all right, all wrong, or mixed. So you know, you kind of get a handle whether they might have guessed. And so they may regroup students, they may use different resources, they may team up with another teacher who might have a better handle on math, or one that has a better handle on literacy, and, you know, kind of switch rooms. There's so many different things they can do.

District leaders developed specific “tools” that both communicated district expectations and provided support to teachers in their use of results from the benchmarks. The Data Analysis Protocol reinforces the benchmarks as a formative assessment by helping teachers to think through the steps of analysis and action as they review the Item Analysis Report. The protocol poses the following questions:

- Using the Item Analysis Report, identify the weakest skills/concepts for your class for this benchmark period.
- How will you group or regroup students based on the information in the necessary item analysis and optional standards mastery reports? (Think about the strongest data and how those concepts were taught.)
- What changes in teaching strategies (and resources) are indicated by your analysis of benchmark reports?
- How will you test for mastery?

District leaders also expected analysis of benchmarks to create an opportunity for teachers to reflect on their instruction. They further reasoned that, in analyzing the benchmarks, teachers could begin to examine their own content knowledge and instructional repertoire with an eye on identifying what professional development and support would be personally beneficial. They expected teachers to use the sixth week of instruction not just to reteach in the same old way, but to find new instructional strategies that would prove more successful. One district administrator described what she hoped would be a teacher’s thought process as she reviewed the benchmarks for her class,

Well, I think the benchmark gives you information about your class, which then will say to you, “OK, I’ve taught inference, and the benchmarks are showing me over and over again the kids aren’t getting inference. *I* need to do something about trying to find a resource for inference.” This takes me back to the desire to have a platform of material that a new teacher or new coach could sit down together and say, this is great, 15 minute, half hour, module, on teaching and assessing the use of inference. But that’s going to take some time.

While this administrator saw that developing the tools and professional development needed to build the kind of capacity needed, other administrators believed that the district had done what it could and it was now in the hands of teachers:

This year we want *everybody* to use the benchmark data. No choices. There *is* a way to use the data and it is specific. Teachers get an item analysis that is specific. How many answered each item and what did it test? Any item where more than half got the answer wrong, you need to reteach it. They got the data protocols in the summer. The principals got trained in a day during the summer. The teachers got trained on the first half day in October. The principals got the PowerPoint and the principals trained the staff. We wrote a script for them.

To encourage teachers’ reflective use of the benchmarks, the district created a single-page “Teacher’s Reflection” protocol intended to be completed by individual teachers following each administration of the assessment. The reflection prompts communicate the expectation that

teachers consider how they are going to instructionally respond to students and to also consider their own instructional and pedagogic needs:

- In order to effectively differentiate (remediate and enrich), I need to...
- Based on patterns in my classes' results, I might need some professional development or support in...

Finally, district leaders expected individual teachers to access and use a variety of analyses of benchmark data available on SchoolNet and to take advantage of instructional features of SchoolNet such as information about how to reteach particular standards, and additional practice worksheets for students.

Expectations for Benchmark Use by Instructional Communities

While the primary focus of central office staff members was on the use of benchmark results by individual teachers, they also expected that various groups in the school—especially grade groups—would examine the data. This expectation that teachers would talk with one another regularly was explained by a district leader who commented:

The expectation is that the 3rd grade teachers will sit at a table with each other and say, “Here’s how my kids did on item 1. How did your kids do? Whoa! My kids didn’t do well. Your kids all nailed it. Tell me how you taught that? Alright, I’ll go back and I’ll try that.” That’s supposed to happen item by item.

When ORS staff piloted the core curriculum and benchmarks in its schools, they communicated the expectation that grade groups should look at benchmark data together, by creating a “standard agenda” for grade group meetings that included reviewing data, making decisions based on data, and then monitoring and following up on the outcomes of those decisions. Once the core curriculum and benchmarks were scaled up across the whole district, this expectation became less explicit and there was much greater variation in whether schools put aside time for grade group teachers to meet. As one administrator explained,

In some schools they do, but not as frequently as they would like to, meet together in grade groups and discuss, “Well, my kids did this on this. Yours did that. What did you do?”

Expectations for Benchmark Use for Whole School Improvement

Beyond the level of instructional communities, district leaders also expected that benchmarks would play a still broader role in whole school improvement. These expectations were largely dependent on strong principal leaders who would reinforce district expectations of teachers and who would create a professional climate that encouraged organizational learning through inquiry, reflection, and informed action. They were also dependent on the alignment and coordination of other school-focused improvement processes including SchoolStat, the School Improvement Planning process, and the work of School Assistance Teams (for low-performing schools).

District leaders realized that teachers’ use of the benchmarks would be limited unless strong, supportive, committed, and engaged leadership was present in schools. More specifically, the district’s expectations of principals included dedicating collaboration time immediately after the benchmark results are reported for teachers to discuss and analyze the benchmarks. One district

leader described the principal's role with regards to the professional climate that would need to be established:

To give teachers the time to have the conversation to plan instruction and to support the teachers in doing what they need to do as far as giving them the resources, the professional development, the climate to feel safe to talk about what they know and what they still need to learn themselves.

The district also expected principals to help with identifying the professional development needs of their faculty, again, based on the results of the benchmarks. Finally, the district expected principals to make certain individual teachers received the professional development they needed.

Alongside their responsibilities within the school, the district expected principals to attend and participate in monthly SchoolStat meetings hosted by their Regional Director. As mentioned earlier, at these meetings, principals, usually grouped by grade levels, would meet to discuss school climate issues and assessment results, including the benchmarks. According to district leaders, the purpose of SchoolStat meetings was to discuss schools' data, note data trends, probe reasons for such trends, exchange best practices and experiences, and develop action plans around data.

Finally, the district expected benchmark data to be used, along with other forms of data, in the annual school improvement planning process. For example, there are repeated references to benchmarks as one form of data to analyze in the required template for school improvement planning provided to elementary and middle schools in February of 2007. There is also a section specifically on benchmarks, which asks questions including: "What do you SEE in the data? What do the data SAY? What QUESTIONS do the data raise? What questions do you have about what you see?" There is also a section for discussion of "Applicable/Relevant Student Groups" (Performance Fact, Inc., 2007, p. 16, emphasis in original). The district focus on benchmarks as formative assessments is reflected in this document, which includes the question: "Is there strong, observable evidence that school staff regularly uses Benchmarks and other *formative assessments* to monitor and adjust instructional practices?" (Performance Fact, Inc., 2007, p. 19, emphasis in original). In the spring of 2007, the district mandated that two professional development days be devoted to the school improvement planning process; for the second of these days, schools were expected to provide analyses of benchmark results for each content area, grade, and test date.

Provider Expectations for Benchmarks

The Office of Restructured Schools was explicit, and some would say, unremitting in its message that its schools would use the core curriculum with fidelity and the benchmarks to guide instruction. This expectation was strong, but support was also forthcoming. As one principal in an ORS school explained,

The District is really making sure that as instructional leaders, we know how to collect the [benchmark] data, how to interpret the data. So we've received a lot of professional development ourselves on how to look at the data, then transfer it into instructional strategies, and tie it into the core curriculum.

While all of the outside providers, except Edison, adopted the district's core curriculum and benchmark assessments, there was initially considerable variation in the degree to which they emphasized the importance of their use. Universal communicated the strongest and clearest expectation that teachers were to look at the data, talk about the data, and make changes in their instruction based upon the data. As the head of Universal's effort explained, "We used benchmarks to death."

The university providers were the most ambivalent about the core curriculum and the benchmarks. It had its own literacy framework and the implementation of the framework was its primary expectation of its partnership schools—at least during the first two years. The other university adopted the core curriculum, but its strongest expectation was that teachers would participate in university-sponsored professional development—some of which was focused on the core curriculum—and would learn about how to construct high-quality, ongoing classroom-based assessments. Over time, leaders in both university partnership schools pressed their partners to shift to an increased emphasis on the core curriculum and benchmarks. There appeared to be two reasons for this shift. The first was that principals and some teacher leaders came to believe that fidelity to the core curriculum and attention to the benchmarks was the most certain pathway to AYP. The second was, as one principal put it, "We have the benchmarks. The benchmarks measure the core curriculum. Therefore, we should be using the benchmarks."

In 2006-07, one nonprofit provider convinced its schools that an additional interim assessment that predicted student performance on the PSSAs would be a nice complement to the district's benchmarks, and perhaps be a stronger assessment strategy for making AYP.

District Supports for Benchmark Assessment Use

District leaders set high expectations for the instructional use of the benchmarks. They acknowledged, however, that these expectations were predicated on having the following things happen: changes in schools' schedules, changing what happens in the classroom, and, to a certain extent, tinkering with long-standing social, cultural, and instructional practices. District leaders were also quick to note a distinction between *ideal* expectations and intended use of the benchmarks and realized practices. That is, they noted the gap between what *could and should be* and what *was* happening on the ground, in individual classrooms and schools.

The District provided four types of supports to *all* schools in the district to try to bring about these changes and to support the intended use of the benchmark assessments: (1) SchoolNet; (2) the Data Analysis Protocol; (3) professional development; and (4) time. The District provided additional supports to low-performing schools.

SchoolNet

The district contracted with SchoolNet Instructional Management Solutions (SchoolNet) to develop a districtwide database for the benchmark assessments and other student data, to make assessment data immediately accessible to every classroom teacher and building principal, and to develop analysis and instructional tools for educators' use. Students' families also have limited access to SchoolNet data through the system's FamilyNet tool to obtain up-to-date information on their children's test scores (including benchmark assessments), report card grades, and attendance.

A critical feature of SchoolNet is the Item Analysis Report. As shown in Figure 3, a mock-up of the report, the report creates data spreadsheets for every teacher that tells her:

- The names of every student who took the benchmark;
- The correct answers for each benchmark item;
- How *many* and exactly *which* items each student answered correctly;
- The *wrong* answer selected by individual students for each item;
- The average percent correct for each class for each item by state standard statement; and
- The state standard statement tested for each item. (SDP website document: Office of Curriculum and Instruction: Planning and Scheduling Timelines, 2007-2008, p. 18.)

Comment [CEL1]: OK?

Students' correct answers are indicated by a green checkmark, while incorrectly answered items are indicated by each student's actual multiple-choice answer (e.g., "A") appearing in the cell in red. The state standard to which each particular test item is linked is noted at the top of the spreadsheet. Additional analysis formats enable teachers to group the data by state standard so they can identify on which standards students did well, or less well. A companion paper in this symposium, describing how Philadelphia 3rd and 5th grade teachers analyze benchmark assessment results and plan instruction based on these results, more thoroughly discusses teachers' use of this analysis tool.

Figure 3: Benchmark Item Analysis Spreadsheet for Philadelphia.

Class-Wide Summary		23 students in this section 20 students took this test																				Total	
Standard ID	Total	1 View	2 View	3 View	4 View	5 View	6 View	7 View	8 View	9 View	10 View	11 View	12 View	13 View	14 View	15 View	16 View	17 View	18 View	19 View	20 View	Total	Standard ID
2.2.5.A.1	20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20	Correct Response
2.2.5.A.1	16/20	A	D	A	C	C	C	B	D	A	A	A	B	D	B	D	B	C	B	C	C	16/20	Point Value
2.2.5.C.1	12/20																					12/20	Summary Score (Points)
2.2.5.A.1	16/20																					16/20	Summary Score (Percent)
2.2.5.C.1	12/20																					12/20	
2.1.5.D.1	19/20																					19/20	
2.1.5.A.1	12/20																					12/20	
2.4.5.A.1	16/20																					16/20	
2.2.5.B.1	20/20																					20/20	
2.1.5.B.1	18/20																					18/20	
2.1.5.E.1	12/20																					12/20	
2.1.5.B.1	13/20																					13/20	
2.4.5.A.1	20/20																					20/20	
2.1.3.I.1	18/20																					18/20	
2.4.5.A.1	18/20																					18/20	
2.2.5.B.1	17/20																					17/20	
2.2.5.C.1	14/20																					14/20	
	327/400																					327/400	Summary Score (Points)
	82%	80%	80%	60%	80%	60%	100%	95%	60%	80%	100%	100%	90%	90%	60%	65%	100%	90%	90%	85%	70%	82%	Summary Score (Percent)

Student-by-Student Data		The list below reveals how each student answered each test item. You can select one or more students to add to a Student Group.																				Total		
Student Name	Total	1 View	2 View	3 View	4 View	5 View	6 View	7 View	8 View	9 View	10 View	11 View	12 View	13 View	14 View	15 View	16 View	17 View	18 View	19 View	20 View	Total	Standard ID	
Abeey Z.	95%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	B	✓	✓	✓	✓	✓	✓	95%	Abeey Z.	
Ananda Y.	50%	D	B	✓	D	B	✓	✓	✓	C	B	✓	✓	D	✓	A	B	✓	✓	✓	✓	50%	Ananda Y.	
Ali X.	70%	✓	✓	C	✓	B	✓	✓	✓	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	C	A	B	70%	Ali X.
Cheyenne W.	90%	✓	✓	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	A	✓	✓	90%	Cheyenne W.
Deiondra V.	95%	✓	B	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	95%	Deiondra V.	
Dakota U.	65%	✓	✓	D	✓	B	✓	C	C	✓	✓	✓	✓	✓	D	B	✓	✓	✓	✓	A	65%	Dakota U.	
Dwayne T.	75%	C	✓	✓	B	✓	✓	✓	✓	B	✓	✓	✓	✓	C	B	✓	✓	✓	✓	✓	75%	Dwayne T.	
Jacy S.	70%	✓	✓	D	✓	B	✓	✓	✓	C	✓	✓	✓	✓	A	B	✓	✓	✓	✓	B	70%	Jacy S.	
Jariah R.	85%	✓	✓	C	✓	B	✓	✓	✓	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	85%	Jariah R.	
Kendis Q.	95%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	C	95%	Kendis Q.	
Lakin P.	90%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	A	C	✓	✓	✓	✓	90%	Lakin P.	
Lenelle O.	50%	C	B	D	D	B	✓	✓	C	B	✓	✓	D	✓	✓	✓	✓	✓	✓	✓	✓	50%	Lenelle O.	
Mekella N.	95%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	B	✓	✓	✓	✓	✓	95%	Mekella N.	
Mancel M.	90%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	A	✓	✓	A	90%	Mancel M.	
Nara L.	100%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100%	Nara L.	
Shandi K.	80%	✓	✓	D	✓	B	✓	✓	✓	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	80%	Shandi K.	
Sidone J.	65%	B	C	✓	B	✓	✓	✓	✓	B	✓	✓	✓	B	A	B	✓	✓	✓	✓	✓	65%	Sidone J.	
Talisa J.	100%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100%	Talisa J.	
Tate H.	100%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100%	Tate H.	
Yancy F.	75%	✓	✓	C	✓	B	✓	✓	✓	A	✓	✓	✓	✓	A	✓	✓	✓	✓	✓	✓	75%	Yancy F.	

SchoolNet provides a number of other on-line features to assist teachers with data analysis and re-teaching, including links to the actual test questions, information about how to reteach the particular standards, and additional practice worksheets for students. To facilitate teachers' use of SchoolNet, the SDP issued laptop computers to all teachers in the district over the course of three to four years. In giving laptops to teachers, the district reinforced its expectation that teachers would access student and classroom data on SchoolNet so they can make informed, "data-driven" instructional decisions.

Principals in the CPRE study schools also accessed SchoolNet to see how students and teachers in their schools were performing. As one principal reported:

I look at it at least once a week, just in case something comes up and I take a look at a specific kid. Or a teacher may be threatening to retain a kid and go back and look at it to see what the problems are. Or you look for the data that's general, like everybody in third grade missed this point. What is it? And you go and you tell them, "All of you missed this one particular skill or this one particular standard. It is either, one, you didn't teach it, or they didn't get it. And you need to go back and do it again....Or when I see one class that is like sort of lagging behind the rest of the classes...then I have to figure out how I can support the teacher (Principal interview, CPRE school #5).

Another principal reported comparing the performance of classes as a way of identifying sources of assistance.

And I'll say [to the teacher with lower-performing students], "Maybe you should talk to Mrs. So-and-So because when she did Time, her kids did a lot better. So maybe she can give you some feedback or strategies that will be successful with the children who didn't (Principal interview, CPRE School #6).

While principals can access data on all of their students and classes, they cannot monitor whether and how often teachers are using SchoolNet. As one principal noted, the District might be looking at how often the teachers log on, but the only way she can be sure teachers are using SchoolNet is through the Data Analysis Protocol that teachers must submit to their principals. Some principals believed that accessing SchoolNet was not a good use of teachers' time. In these cases, either they or their school's teacher leaders organized and printed the data for teachers. One principal explained,

You know the district can now check and see how many teachers in my school are logging on to SchoolNet. This is ridiculous because I don't think that's a wise use of my teachers' time (Principal interview, RFA).

Data Analysis Protocol and Other Tools

As discussed in the section on District Expectations, the District designed a Data Analysis Protocol to assist teachers in evaluating data from SchoolNet's Item Analysis Report. Teachers were to use this report to identify the weakest skills and concepts for their class in order

to plan lessons during the reteaching weeks, to group or regroup students, and to examine their strategies and make mid-course adjustments. Teachers could also use information on individual students to provide targeted support to students based on their needs. Finally, the Teacher Reflection sheet enabled teachers to identify professional development that would be of help to them. It is notable that neither of these protocols mentions or refers teachers to the district's core curriculum, despite the fact that the curriculum offers many resources and strategies for teachers to use to differentiate instruction.

Teachers are required to submit the Data Analysis Protocol Worksheet to their principals along with a copy of the Item Analysis from SchoolNet. Principals in the CPRE study schools expected teachers to discuss their reports in their grade group meetings and they used their own review of the Protocols as another opportunity to give feedback to their teachers. The principals, in turn, completed Protocols on their schools and reviewed these reports with their Regional Superintendents.

District leaders developed additional tools for teachers and school leaders as they began to recognize the need for other kinds of support. For example, one administrator in the Curriculum Office identified a need, while also insisting that central office was playing a guidance role rather than a prescriptive role: "We realized that teachers are not using the sixth week well so we wrote suggestions for them this summer for reteaching. There are choices. Most teachers are glad to have some guidance. We are *not* scripted. What we have is instructional guidance." Again, we see that District-developed supports also communicated expectations to teachers about the use of benchmarks.

Professional Development

The District provided several kinds of professional development for teachers: (1) on the district curriculum; (2) on the use of SchoolNet; and (3) school-based coaches. In addition, several principals in the CPRE study schools reported that the Regional SchoolStat meetings and follow-up activities with other principals provided another source of professional development for them and some of their staff.

When the district implemented its new mathematics curriculum, Everyday Mathematics (EDM), all teachers attended district-run multiday training sessions in the summer prior to the enactment of the curriculum. New teachers also have the opportunity to attend EDM training as well. The district expected all teachers to receive training on the use of SchoolNet, but used a school-based, turnkey training approach. Generally, principals and a technology support person received professional development from the central office and were expected to return to their schools and train their staff. The principals interviewed in the CPRE study said they expected all of their teachers to learn how to use SchoolNet, either through school-based professional development (often offered after school or on weekends) or on their own. The technology support person was usually on call to assist teachers in the building.

Buildings also had school-based literacy and, often, math coaches. The number and mix of coaches sometimes depended on availability of funding. Principals might choose, for example, to use their professional development funds for coaches. In the CPRE schools, the literacy

coaches tended to be full-time coaches, while math coaches were teachers with some (or no) released time. In the RFA schools, there was a range from full-time to no released time for coaches in both areas.

Principals attended meetings with their Regional Superintendents and other principals in their region on a regular basis. Some of these meetings were devoted to reviewing the SchoolStat data and discussing among themselves ways of addressing issues of student behavior and performance. The principals in the regions created other opportunities to share their practices, however. One group of principals, for example, brought teachers and math and literacy coaches together by grade level to share “best practices.” One principal explained:

X School were the specialists of grade two. All the grade two teachers from the schools [in their cluster] reported to X school. They presented. They exchanged best practices. They came back to school the next morning and they couldn't stop raving about just sharing and talking to another second grade teacher [from another school]...They came back with packets to provide—they did reflective, turn around training (Principal interview, School #).

The resulting product was a “best practices” binder with a section contributed by each school that principals felt would provide teachers with new instructional ideas.

Time

Using the benchmark assessment results for instructional purposes requires time: to re-teach skills, to analyze data, and to partake in professional development.

The primary mechanism to support teachers' use of the benchmark assessments is the reteaching week scheduled after each benchmark assessment. As discussed in earlier sections of the paper, the assessments are electronically scored and, in a quick turnaround, individual and class results are made available to teachers through SchoolNet. During the remaining five days of the cycle, the sixth week, the teachers are expected to plan and execute their reteaching, remediation, and enrichment activities. While the Data Analysis Protocol asks teachers how they will group or regroup students based on their analysis of the assessment results, the teachers enjoy considerable latitude around the specifics of instruction and reteaching. Students are not retested on the content of the instructional cycle. After the reteaching week ends, students move on to the next instructional unit.

Both district staff and principals in the CPRE study schools expected teachers to meet in grade groups to discuss their students' performances on the benchmarks and to share instructional strategies and concerns with one another. To expedite this sharing, elementary school teachers in the same grade were given a common planning time. In addition, the district instituted “half-day Fridays” every other week. On these particular days, students are released around noon and teachers remain in the building for professional development workshops and sessions. As one central office leader explained:

The Chief Academic Office was very focused on what do we need to do to support the teachers, for them to use the [SchoolNet] system? And so, one of the first things is getting mandated days on the school district calendar so that we

know that in every single school, people will be looking at the same thing, and learning the same thing.

It was up to the individual principals, however, to ensure that the allotted time was used to analyze and discuss student results and to learn about new instructional techniques.

Supports for Low-Performing Schools

The School District of Philadelphia has implemented several additional measures to monitor and support low-performing schools that are not meeting their AYP targets. A School Assistance Team (SAT) is assigned to every school that is in Corrective Action. A SAT is headed by a case manager who may be a retired or current district administrator and has several additional team members. The team is responsible for collecting data about a school, analyzing the data, and writing a report that is shared with the school. This data includes observations in all classrooms, interviews with staff, students and parents as well as other kinds of data about student achievement, attendance, school climate, etc. Working with its case manager, the school leadership team is responsible for developing a plan to address issues raised in the SAT's report. The case manager and an Intervention Administrator work with the school to implement the plan and monitor progress. For example, the SAT report might say: "While the school leadership team has a strong understanding of student achievement based on a variety of kinds of data, this understanding does not permeate the entire staff and teachers appear to be much less knowledgeable about what assessment data indicate about their students' learning." A second example might be: "Observations of math lessons in grades 4, 5, and 6 indicate that teachers are spending lots of time on computation skills at the expense of problem solving skills and mathematical concepts. This suggests that teachers may not be comfortable with the Everyday Math curriculum."

In 2006-07, the district created the position of School Growth Coordinator (SGC) for the 129 schools that were in "School Improvement" status. Schools that were managed by an outside provider were exempt from the requirement, although all of the RFA schools opted to have the position. School Growth Coordinators are responsible for organizing and analyzing data about student performance and working with teachers to understand the data and develop instructional interventions based on the data. They work closely with their principals to develop and implement the School Improvement Plan.

Challenges to Meeting District Expectations for Benchmark Use

Accountability Issues

Our analyses of interviews with district leaders and review of district documents indicate that the primary intended purpose of the Benchmarks was to inform classroom instruction. However, district leaders' comments revealed summative purposes as well. As one district staff person commented in 2003, "In the long term we looked at what were the increases we wanted to make and then we said, how would we check them along the way. So, we established at that time that we would have benchmarks given every six to eight weeks that would simulate the state tests." But then she went on to say, consistent with the emphasis on formative assessment, that "we would have the teachers receive that information back and that would be used to inform instruction." Also, in acknowledging that the benchmarks, in effect, created a pacing cycle that

principals could use to monitor teachers' implementation of the core curriculum, district leaders recognized a summative purpose for benchmarks.

The use of benchmarks as part of the SchoolStat process and the School Assistance Team process heightened school-based leaders' perception that benchmarks were summative and, indeed, part of the district's accountability system. In these settings discussion of benchmark results occurred in settings where administrators from central and regional offices—some of whom had line and rating authority for principals—were present. During SchoolStat meetings focused on student performance, benchmark data were aggregated at the school level and reported by grade, subject and state standard, showing for each school in the region the percentage of students who scored 75% or above, between 50% and 74%, and below 50%. One regional superintendent explained how she saw SchoolStat changing the purpose of the benchmark assessments:

This past year with SchoolStat, it now became summative. And all of a sudden this formative data become summative, and for me, it sort of lost the essence of how do we improve practice in the classroom? And I don't believe that...these benchmarks are a summative thing. I mean, nobody said that they were high stakes test, yet we're treating them as high stakes tests.

However, a principal in another region described SchoolStat more as a support than an accountability mechanism, "The elementary principals sit in the room and she [regional superintendent] puts the data up there in full view of every school. Not comparing but everybody sees it and then we go around the table and talk about why it's better, or why it's worse." These divergences reinforce Halverson et al.'s (2007) point that the distinction between formative and summative is in the eye of the beholder.

Alignment Issues

Most of our school and district respondents were satisfied with the degree of alignment between the benchmark assessments, the Pennsylvania state standards, the state assessment anchors, and the elementary curriculum. The district raised a few issues around alignment, however.

First, one district administrator felt that the benchmark assessments in mathematicsⁱⁱⁱ are easier than the state test. While she agreed that the benchmarks were conceptually aligned with the state test, she argued that

...the level of questions—when I look at the level of [PSSA] questions on the website... When you look at those sample questions, one thing that strikes you...is the amount of reading, even in math. And there are a few sentences, three sentences, four sentences that the kids have to read carefully and understand in many of the math PSSA problems. But in the very same concept, we test [on the benchmarks] by reducing the amount of reading. So, although we must say that it's the same concept, but it's not the same. Because when those kids are in the benchmark test are not used to as much reading, and they just read two sentences and they get the information to do it. And [on] the PSSA, they have to read four sentences and very carefully. It makes a difference.

Second, district respondents expressed concern that the benchmarks do not contain open-ended items, a format that is used on both the mathematics and writing portions of the state test.

As a result, Philadelphia looked for a provider to supplement the benchmarks with these kinds of items, but ran up against time issues. As a curriculum administrator explained:

It's [the benchmark assessment] not in an open-ended, constructive response format because despite the fact that people will tell you they are going to be able to do that by the second test of the year, so hire us, they don't have anybody who can do a timely turnaround of results.

Third, a district leader in the Office of Curriculum and Instruction describes issues that are raised when the format of the state test is changed:

Because the PSSA used to have seven items that were noncalculator items, last year [2004-05] we had five noncalculator items. You do your noncalculator items, fold the page, and then go on and get your calculator if you need it and do the rest of the test. We looked at that and we said we're still weighting the test—five out of 20 items is weighting the test far more heavily to noncalculator items than the PSSA does. So this year [2005-06], we went with four noncalculator items. The state has also changed. Now they only do four noncalculator items. So we're trying to figure out what to do next year. Should we just blow off the noncalculator items? Should we do two noncalculator items? Is it worth having two noncalculator items? I don't know. Maybe. Obviously that would be 10% of the test. It's only about 5% of the PSSA. So that's how as things change, we move along with what the state is doing.

Technical Issues

There has been growing tension in the district about on-line test administration of the Benchmarks. In the initial years, students took the test on paper and then typically, the teacher received some time at a computer lab for the students to enter their answers for each question into the computer-based test. In 2006, however, the district evaluation unit moved to have students complete the tests online to save time. Curriculum Office, in turn, has fought to maintain the option of having students take the assessments in pencil-and-paper format as its leaders believe that the computer format was poorly executed and required extensive scrolling up and down to see all of the components of each question. For the 2006-07 school year, 3rd graders received a dispensation from this online test administration policy. Further complicating this dilemma is that some teachers like having the written tests so they can see student work on the booklets and therefore the kinds of mistakes that their students make on the test. In this way, the hard copies of students' benchmarks provide teachers with an often-rich "paper trail" of their students' thinking. As well, the tension around how students take the benchmarks spills over into the instructional challenges.

Instructional Issues

While the benchmark assessment system provided periodic information on students' performance on the district's core curriculum and the management information system facilitated analysis of these data by teachers and their principals, they did not address the question of "What should I do next?" District leaders expected teachers to analyze data in grade groups and to develop and/or share appropriate instructional practices to meet the needs of their students among themselves with support from their teacher leaders, coaches and principals. The core

curriculum offered extensive resources and alternative teaching strategies for helping students who experience difficulty. SchoolNet provided information about how to reteach particular standards and additional practice worksheets for students. However, as we will see in the papers that follow in this session, these resources were used much more infrequently than district leaders anticipated.

Two additional instructional issues were named by several interviewees. The first relates specifically to the district's math curriculum, Everyday Mathematics. This curriculum is a challenging one for teachers because it spirals, returning to concepts previously taught over and over again, each time developing the concept more deeply. A district administrator explained, "The biggest headache for teachers is that they feel uncomfortable going on before the kids have mastered certain things." The second relates to all of the subject areas. The benchmark items are at grade level. If a student reads significantly below grade level—as do many students in Philadelphia—the benchmarks do not provide information about that student that will be particularly helpful to a teacher.

Resource Issues

As the district has made increasing cuts in its budget due to inadequate funding, resources in support of the use of benchmarks have evaporated. For example, the District cancelled the SchoolStat contract. (Some regional superintendents have tried to maintain these data meetings. In one of the regions included in the CPRE study, the regional director reported holding such meetings focused on attendance, suspensions, and specific standards across schools in the region. There has also been some talk at the regional and district level that the district might hire college students to help the different regions prepare data for SchoolStat-like meetings.) In addition, the bimonthly school-level professional development sessions have been abolished. Finally, as school-based budgets have been reduced, principals have had to eliminate teacher leaders or cut back on their release time, making it more and more difficult for these leaders to meet with teachers about their data and suggest new instructional strategies.

Conclusion

In the past six years, the Philadelphia school district has undergone dramatic, and at times tumultuous, changes. These changes have included a more prescriptive approach from the district about what should be taught and when it should be taught, as well as new strategies to provide assistance to schools and teachers. Benchmark assessments have been, from the district perspective, an important link between district policies and instructional practices. The hope has been that the assessments themselves, along with the supports and requirements that accompany them, will serve a formative function for teachers as they seek to improve student learning. The six-week cycle of benchmarks in grades 3-8, with the sixth week devoted to reteaching areas identified as weak through the assessments, was designed by district administrators with the expectation that this additional time would enhance the formative uses of the assessment. It is our assessment that the benchmarks can serve a formative purpose, that the six-week instructional cycle enhances their formative use, and that the tools that the district has provided is a useful set of supports and tools to help school-based educators use them formatively. In this paper, we have outlined district expectations and supports. As we will see in the papers that follow, teachers and school leaders are, in general, positive about the core curriculum and benchmark assessment. Nevertheless, they have encountered significant obstacles—some

anticipated, some not—in using them formatively; that is, to understand what their students know and can do and to address weaknesses.

Our literature review highlighted the distinction between summative, interim, and formative assessments, including the ways in which these distinctions are based not only on the design of the assessments, but also on how they are used and *experienced*. The Philadelphia case demonstrates, among other things, that even formative assessments can become summative in settings in which people are differently positioned *vis a vis* power and authority and in an overall environment of high-stakes accountability, .

References

- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–75.
- Boyd, W. L., & Christman, J. B. (2003). A tall order for Philadelphia's new approach to school governance: Heal the political rifts, close the budget gap, and improve the schools. In L. Cuban & M. Usdan (Eds.), *Powerful reforms with shallow roots: Improving America's urban schools* (pp. 96-124). New York City: Teachers College Press.
- Bulkley, K. E., Mundell, L. M., & Riffer, M. (2004). *Contracting out schools: The first year of the Philadelphia Diverse Provider Model* (Research Brief). Philadelphia: Research for Action.
- Burch, P. E. (2005). *The new educational privatization: Educational contracting and high stakes accountability*. Retrieved December 20, 2005, from <http://www.tcrecord.org/content.asp?contentid=12259>
- Christman, J. B., Gold, E., & Herold, B. (2005). *Privatization "Philly Style": What can be learned from Philadelphia's diverse provider model of school management?* (Research Brief). Philadelphia: Research for Action.
- Halverson, R., Pritchett, R. B. & Watson, J. G. (2007). *Formative feedback systems and the new instructional leadership*. WCER Working Paper No. 2007-3. Madison, WI: Wisconsin Center for Education Research.
- Hill, P. T., Campbell, C., & Harvey, J. (2000). *It takes a city: Getting serious about urban school reform*. Washington, DC: Brookings Institution Press.
- Hill, P. T., Pierce, L., & Guthrie, J. (1997). *Reinventing Public Education : How Contracting Can Transform America's Schools (RAND Research Study)*. Chicago: University of Chicago Press.
- Maranto, R. (2005). A tale of two cities: School privatization in Philadelphia and Chester. *American Journal of Education*, 111(2), 151-190.
- Olson, L. (2005, November 30). Benchmark assessments offer regular checkups on student achievement. *Education Week*, pp. 13-14.
- Performance Fact, Inc. 2007. *Getting results! 2007 to 2008 Continuous School Improvement Plan, Elementary and Middle School Template. Customized by Performance Fact, Inc. for the School District of Philadelphia*. Oakland, CA: Author.
- Perie, M., Marion, S., & B. Gong, & Wurtzel, J.. (2007). *The role of interim assessments in a comprehensive assessment system: A policy brief*. Washington, DC: The Aspen Institute. A framework for considering interim assessments. National Center for the Improvement of Educational Assessment. Unpublished Manuscript.
- Research for Action. (2005). *The "Original 86."* Philadelphia: Research for Action.
- School District of Philadelphia. (September, 2007). *Benchmark Assessments 2006-2007, Grades 3-8*. Retrieved from: www.phila.k12.pa.us/offices/curriculum/webinars .

- Travers, E. (2003, November). *The state takeover in Philadelphia: Where we are and how we got here*. Retrieved May 1, 2006, from <http://www.researchforaction.org/PSR/PublishedWorks/ST031004.pdf>
- Useem, E. (2005). *Learning from Philadelphia's school reform: What do the research findings show so far?* (Occasional Paper). Philadelphia: Research for Action.
- Wong, K. K., & Shen, F. X. (2003). Measuring the effectiveness of city and state takeover as a school reform strategy. *Peabody Journal of Education*, 78(4), 89-119.

ⁱ This section is a substantially revised and updated version of an earlier report by Useem, Christman and Boyd, *The Role of District-level Leadership in Radical Reform: Philadelphia's Experience Under the State Takeover, 2001-2006*. Philadelphia: Research for Action and Temple University, Mid-Atlantic Educational Laboratory for Student Success, College of Education. <http://www.researchforaction.org/publication/details/238>

ⁱⁱ Brian Gill et al., *State Takeover*.

ⁱⁱⁱ The CPRE study focused on mathematics curriculum, instruction, and assessment.