SECTION 2: CASE STUDY: BENCHMARKS IN THE SCHOOL DISTRICT OF PHILADELPHIA

Building With Benchmarks: The Role of the District in Philadelphia’s Benchmark Assessment System

Katrina E. Bulkley
Montclair State University

Jolley Bruce Christman
Research for Action

Margaret E. Goertz
University of Pennsylvania

Nancy R. Lawrence
Math and Science Partnership of Greater Philadelphia

In recent years, interim assessments have become an increasingly popular tool in districts seeking to improve student learning and achievement. Philadelphia has been at the forefront of this change, implementing a set of Benchmark assessments aligned with its Core Curriculum district-wide in 2004. In this article, we examine the overall context for Benchmarks in Philadelphia, the expectations district leaders had for the use of those Benchmarks, the supports put in place to assist those in schools in meeting those expectations, and the challenges encountered in that implementation.

Research for this article was supported by a National Science Foundation Grant REC-0529485 to the Consortium for Policy Research in Education 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.

Correspondence should be sent to Katrina E. Bulkley, College of Education and Human Services, Montclair State University, 1 Normal Avenue, Montclair, NJ 07042. E-mail: Bulkleyk@mail.montclair.edu
In recent years, as the push for increased student achievement has intensified, districts are using “interim assessments” to track students’ progress at regular intervals throughout the year. Although some districts are constructing 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, score them, and analyze results (Burch, 2005). These new assessments, often called “benchmark assessments,” are given multiple times a year in various subjects (see Bulkley, Nabors Oláh, and Blanc, 2010/this issue, for a description of “interim assessments”). According to 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 adequate yearly progress (AYP) target. However, some districts have implemented interim assessments that are more closely aligned to a district-wide curriculum or district standards and are intended to provide educators with formative information about students’ mastery of the curriculum (Olson, 2005).

Perie, Marion, Gong, and Wurtzel (2007) discussed three core purposes for interim assessments: instructional, evaluative, and predictive. Instructional purposes “provide results that enable educators to adapt instruction and curriculum to better meet student needs” (p. 4). Examples of such purposes include enriching the curriculum, determining students’ strengths and weaknesses, and providing feedback for students. Interim assessments that serve an evaluative purpose provide information designed to allow educators at the school or district level to make changes at the programmatic level to improve instruction and, ultimately, student performance. Such evaluative purposes can also serve to “enforce some minimal quality through standardization of curriculum and pacing guides” (p. 5). Finally, predictive purposes for an interim assessment are primarily that they “are designed to determine each student’s likelihood of meeting some criterion score on the end-of-year tests” (p. 5).

Despite the increasing use of benchmarks, we know little about the expectations district staff have for these tests. We also know little about the kinds of policies and supports that might facilitate the use of benchmarks for instructional purposes at the level of individual students, classrooms, or schools. Research reported elsewhere describes the central role that benchmark assessments have played in Philadelphia as well as the overall satisfaction of teachers with these assessments (Christman et al., 2009). This combination of relatively early adoption of this form of interim assessment and strong evidence about its use makes Philadelphia an excellent case study for exploring these issues. In this article, we address these research questions:

1. What are the accountability, organizational, and instructional contexts for Benchmark assessments in Philadelphia?
2. What are district leaders’ expectations concerning the use of the Benchmark assessments?
3. What supports do schools receive for use of the Benchmark assessments and instructional improvement?
4. What are the challenges to meeting district leaders’ expectations for use of the Benchmark assessments to improve instruction?

We begin the article with a discussion of the policy context for the development of the Benchmark assessments in Philadelphia. The second section presents the research methods for two studies on which this article draws, whereas the third section describes the assessments themselves. The fourth and fifth sections present our findings on the district’s expectations for
the use of the assessments and district supports for their use, respectively. The article concludes with challenges to meeting the district’s expectations for the use of the Benchmarks.

THE CONTEXT FOR BENCHMARK ASSESSMENTS IN PHILADELPHIA

Since a state takeover in 2001, the School District of Philadelphia (SDP) 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. At the same time, it instituted strong centralizing measures including a district-wide Core Curriculum, mandated after-school programs, and conversion of middle schools to K–8 schools. The district also became a national frontrunner in welcoming the spirit and accountability mechanisms of the 2001 federal No Child Left Behind Act, and low student performance on the Pennsylvania System of School Assessment (PSSA) played an important role in the state takeover and district assignment of a number of schools to outside providers. The SDP has combined what Wong and Shen (2003) described as the leading—and many would say contradictory—alternatives for reform strategies, namely, market-based solutions along with a strong centralized authority model.

Curriculum, Assessment, and Accountability

District-wide assessment has been an integral part of Philadelphia’s education reforms for over 25 years (Boyd & Christman, 2003; Porter, Chester, & Schlesinger 2004). During this time, test results have been used for both accountability and instructional purposes. During 2001, issues around student assessment also gained public prominence in Philadelphia as the state moved toward a takeover of the district, following tensions around both fiscal equity and student achievement. Following negotiations, the city and state compromised on a “friendly takeover” in December 2001 (Boyd & Christman, 2003; Maranto, 2005). Six months later, the newly created School Reform Commission hired former Chicago Public Schools CEO Paul Vallas to head the Philadelphia district. One of Vallas’s first initiatives was to institute a district-wide Core Curriculum in four academic subjects for grades K–8 that was mandatory for regular district schools but voluntary for those managed by providers. Increased testing, including the 6-week formative Benchmark tests, also accompanied the new Core Curriculum (Useem, 2005).

In addition, the district introduced two new performance and data management systems, SchoolNet and SchoolStat, which were rolled out over a period of years. 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, teacher leaders, and principals. SchoolNet is a data management system that provides a variety of analytic reports about student achievement, attendance,

---

1This section is a substantially revised and updated version of an earlier report by Useem, Christman, and Boyd (2006).
and behavior and that also provides instructional tools and information to teachers, principals and families. Schools came online to SchoolNet in cohorts over $2\frac{1}{2}$ years, with those schools managed by outside providers participating in the last cohort.

SchoolStat, on the other hand, was an initiative launched by the district and the University of Pennsylvania’s Fels Institute of Government. It compiled, compared, and presented school-level data on student performance, including Benchmark data, student behavior and student and teacher attendance. SchoolStat was rolled out to 15 elementary schools during the 2003–04 school year and was expanded district-wide by late spring of 2005. SchoolStat was used at regular meetings of regional superintendents with their principals and at meetings of the regional superintendents with the district’s chief academic officer, to discuss the status of, and ways to improve, climate and achievement in their schools. The SchoolStat contract was canceled in 2007, a casualty of budget cuts. Continuing these meetings was left to the discretion of the regional superintendents.

The Diverse Provider Model

Although not initiated by Vallas, the creation of a diverse provider model was probably the most publicized change to result from the state takeover of the SDP. The diverse provider model was a response to a push from the state to create a more “market-based approach to the challenges facing Philadelphia public schools” (Bulkley, Mundell, & Riffer, 2004, p. 1; see also Hill, Pierce, & Guthrie, 1997). In total, seven different organizations (three for-profit educational management organizations, two locally based nonprofits, and two universities) were hired and given additional funds to provide some level of management services in 46 of the district’s 264 schools (Bulkley et al., 2004). The School Reform Commission 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). All of the schools were selected in part due to low performance on the PSSA.

In theory, a diverse provider model involves shifting decision making away from districts and toward 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 cited 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.

METHODS

The findings reported here are drawn from data collected for two separate studies in Philadelphia by researchers from the Consortium for Policy Research in Education (CPRE) and Research for Action (RFA) between 2005 and 2007; the data for each study are described next. Data from each study were coded separately using Atlas.ti and then relevant data and summaries were analyzed collaboratively by researchers from each project.

There are two important issues to note about the data used for this article. First, although the schools in both studies served high-poverty, predominantly minority students, the CPRE study
focused on high-performing schools, whereas the sample for the RFA study included only low-performing schools. Second, the CPRE study focused on the use of the Philadelphia Benchmarks in mathematics, whereas the RFA study included both mathematics and literacy. In the following sections, we distinguish findings, when appropriate, between these two samples.

CPRE Study—Using Formative Assessments: The Role of Policy Supports

The CPRE study focused on mathematics curriculum, instruction, and assessment. In 2006, CPRE interviewed five SDP leaders, including regional superintendents and curriculum, assessment, and technology leaders from the Central Office about the Benchmark assessment system, the district’s expectations for the Benchmarks, the mathematics curriculum (Everyday Mathematics), data analysis and use, and professional development. As well, CPRE interviewed the principals of the six participating elementary schools in 2006 and five of these principals again in 2007. CPRE’s school sample included six schools that had made AYP the year prior to the study and had scored at or above district average on the mathematics portion of the PSSA. These schools mirrored the demographics of the district. All schools were Title I schools, and more than 90% of students in each school were either African American or Latino. A central purpose of the principal interviews was to identify district- and school-level expectations for Benchmark assessment use and potential supports offered to teachers.

In addition, CPRE researchers attended professional development sessions, technology training sessions, and principal meetings (SchoolStat meetings), during which several types of “key 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.

RFA Study: Learning to Learn From Data

Included in the analysis for this article were interviews conducted during the 2005–06 school year by RFA researchers of five SDP administrators from the offices of Accountability, Assessment, and Intervention; Curriculum; and Professional Development. The topics covered included the Core Curriculum, assessments of student achievement, 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 different providers operating in the district were making use of the district’s Core Curriculum and Benchmark system.

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, 2 schools were part of the district’s homegrown intervention under the Office of Restructured Schools, and 1 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 school 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 the RFA schools served a considerably higher percentage of students living in poverty than the district average and were predominantly either African American or Latino.
Between 2005 and 2007, RFA conducted interviews with the principals and teacher leaders in these schools. Each principal was interviewed two to three times and each teacher leader one to two times. Interviews 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. In addition, RFA researchers collected school documents and observed leadership team meetings and grade group meetings where Benchmark results were discussed and professional development sessions on the use of Benchmark data. Researchers also interviewed staff from the education provider organizations to understand the policies and supports related to data use offered by these organizations to the schools that they were managing.

THE PHILADELPHIA ELEMENTARY AND MIDDLE GRADES
BENCHMARK TESTS

Benchmark assessments were implemented district-wide in grades K-8 in Philadelphia in October 2004. Prior to that, they had been used in the set of schools managed by the district’s ORS as part of the diverse provider model. The SDP used Benchmark assessments in Grades 3 to 8 to give teachers immediate feedback relative to their students’ mastery of the Core Curriculum on topics taught in a series of cycles. As shown in Figure 1, each cycle of instruction and assessment consisted of 6 weeks: 5 weeks of instruction at the end of which the Benchmark assessments were administered and a 6th week of review and/or extended development of topics.\footnote{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

---

**FIGURE 1** District theory of change for Benchmark use.

---

[Diagram of School Cycle, Instructional Community Cycle, and Teacher Cycle with specific weeks and activities labeled, such as Weeks 1-4 Core Curriculum Taught Following Pacing Guide, Week 5 Students Take Benchmarks, Teachers Receive Benchmark Results Via Schoolnet, Teachers Analyze Schoolnet Results & Decide Upon Instructional Changes, and Week 6 Instructional Changes Implemented.]
implement an instructional response and test for mastery. As the two articles about data use at the school and individual level suggest, there was extensive variation in how these steps were enacted across schools (see Blanc, Christman, Liu, Mitchell, & Travers, 2010/this issue, and Nabors Oláh, Lawrence, & Riggan, 2010/this issue). The Philadelphia Benchmarks were consistent with Perie and colleagues’ (2007) definition of an interim assessment, in that they “(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” (p. 4).

In Philadelphia, each Benchmark assessment was designed to test only those concepts and objectives taught since the last Benchmark. All of the items in the Benchmark assessments were multiple choice and were designed to assess concepts and skills included in the District’s pacing guide (called the “Planning and Scheduling Timeline”). At the time of the study, the district administered Benchmarks in Reading and Mathematics to students in Grades 3 to 8, and the assessments were cocreated by the SDP curriculum and assessment administrators and Princeton Review in the weeks or months prior to their administration. According to central office administrators, this process required several iterations for each of the five assessments given throughout the school year. The assessments were designed to be aligned to Pennsylvania’s assessment anchors (and, therefore, to the content of the PSSA) and to the district’s Core Curriculum and the state standards. According to one district insider who was interviewed in 2005,

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 [Grades] 3 to 8... into six-week chunks.

Because the Philadelphia Core Curriculum included content that was not tested on the PSSA, the Planning and Scheduling Timelines also identified which concepts were eligible content on the state test and identified the most important “Test Connections” for each unit.

DISTRICT EXPECTATIONS FOR BENCHMARK ASSESSMENT USE

On the district’s Web site, the Office of Curriculum and Instruction 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
- To provide data to assist in instructional decision making

districts in either Pennsylvania or across the nation where time was explicitly set aside for addressing weaknesses found through analyzing benchmark data.
Although the district’s web site formally identified these purposes for the Benchmarks, analysis of interviews with central office staff suggested two central goals. First, the Benchmarks would provide feedback to teachers about their students’ success in mastering concepts and skills covered in the Core Curriculum during each 5-week instructional period. Second, the 6-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 more institutionalized (see Oláh et al., 2010/this issue).

Overall, these purposes reflect a primarily instructional function for Philadelphia’s Benchmark assessments. However, as discussed next, both evaluative and predictive purposes also came into play, with high- and low-performing schools experiencing somewhat different district expectations. The explicit link between Benchmarks and PSSA practice reflect a predictive aspect of the assessments; however, it is worth noting that this “purpose” was not included in district descriptions of the Benchmarks until 2005.

Central office leaders emphasized the use of Benchmark data by individual teachers to understand what their students were learning, but 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. Formal instructional communities are groups of educators who meet regularly to discuss issues of instruction (see Blanc et al., 2010/this issue, for more discussion of instructional communities). The nested nature of these cycles is shown in Figure 1. It was not always clear, however, 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”—that is, tools that would provide support to teachers’ instruction through timely information about what their students were and were not learning. This corresponds to “instructional purposes” in the framework of Perie et al. (2007) and is closest to conventional definitions of formative assessments. When asked, “What would it [use of Benchmark data] look like hands-on for teachers in the ideal?” one district leader responded,

Well we’ve asked them “how might you re-group 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 re-group 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. SchoolNet reports were designed to provide teachers a user-friendly way to access and manage data from Benchmark assessments. The Item Analysis Report, for example, generated data spreadsheets for every
teacher that included the correct and incorrect answers each student had selected, how many and exactly which items each student answered correctly, the average percentage correct for each class for each item by state standard, and the state standard tested for each item. A mock-up of the report can be found in Nabor Oláh et al. (2010/this issue).

A Benchmark Data Analysis Protocol, which teachers were strongly encouraged to hand in to principals, reinforced the expectation that Benchmarks would be used for instructional purposes by helping teachers to think through the steps of analysis and action as they reviewed the Item Analysis Report (discussed at greater length in Nabor Oláh et al., 2010/this issue). The protocol posed the following prompts:

- 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 hoped that the analysis of Benchmarks would also 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 6th week of instruction not just to reteach in the same 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, “Okay, 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.”

Teachers were also asked (but not required) to complete a single-page “Teacher’s Reflection” protocol along with their Benchmark Data Analysis Protocol. This protocol also encouraged teachers to reflect on their instructional practices and prompted them to identify their professional development needs in response to the following:

- 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 . . .

These questions suggest an evaluative purpose for the Benchmarks as teachers (and, potentially, school leaders) were expected to identify what they did not understand in terms of content knowledge.3

3Discussions of the Teacher Reflection Protocol were much more common in the low-performing schools studied by RFA.
Expectations for Benchmark Use by Instructional Communities

Although 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 interest in having groups of teachers regularly discuss the Benchmarks and their students’ performances was consistent with an emphasis on Benchmarks serving instructional purposes. As one district leader explained,

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 the Office of Curriculum and Instruction piloted the Core Curriculum and Benchmarks in the District’s Restructured schools, ORS coaches assigned to each school worked closely with the school’s leadership team and communicated the expectation that time should be set aside for grade groups to meet and look at Benchmark data together. ORS also created 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, however, this expectation became less explicit and there was much greater variation in whether schools put aside time for grade group teachers to meet.

Expectations for Benchmark Use for Whole School Improvement

District leaders expressed more “evaluative” and “predictive” purposes for Benchmark data at the school level, especially for low-performing schools. District leaders expected that Benchmarks could play a broad role in whole school improvement for low-performing schools, thus helping schools to meet AYP. These expectations were largely dependent on strong principal leaders who would be able to understand and communicate the district’s expectations of teachers and who would create a professional climate that encouraged organizational learning through inquiry, reflection, and informed action. 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, as a whole and as individual teachers, based on the results of the Benchmarks. At the same time, the district conveyed to low-performing schools a predictive purpose for Benchmarks, in that Benchmarks were to be used as part of a school’s strategy for making
AYP, particularly in identifying, supporting, and monitoring students who were close to reaching proficiency.4

Alongside their responsibilities within the school, the district expected principals to attend and participate in monthly SchoolStat meetings hosted by their Regional Superintendent. 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.

The use of Benchmark and other forms of data for school improvement (and meeting AYP) was also dependent on the alignment and coordination of other school-focused improvement processes including SchoolStat, the work of School Assistance Teams (described below), and the district’s formal school improvement planning process. In the spring of 2007, the district mandated that 2 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

For schools operated by an outside provider, the expectations of the provider around Benchmarks were relevant alongside the district expectations. Although 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.5 Universal, a community-based organization, 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. One university had its own literacy framework and the implementation of the framework was its primary expectation of its partnership schools—at least during the first 2 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 making 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.”

---

4 Although using the Benchmark assessments for predictive purposes has a certain logic, given that the district believed the Benchmark Assessments to be aligned with state standards, we were unable to find any evidence that prediction studies had been done in data collection for either the CPRE or the RFA studies.
5 Edison used their own Benchmark Assessments, which were part of the overall instructional program they brought to the schools they managed.
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 the following conditions: changes in schools’ schedules to allow for reteaching and regular bimonthly professional development; 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 intended use of the Benchmarks and realized practices. District leaders acknowledged that although progress had been made, there was still a 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: (a) access to online data, resources, and reports through SchoolNet; (b) a structured format for analyzing Benchmark data with the Benchmark Data Analysis Protocol; (c) professional development; and (d) time within the instructional cycle. The district provided additional supports to low-performing schools. All of these are discussed next.

SchoolNet and Related Tools

The Benchmark Data Analysis and Teacher Reflection protocols just described, tied to the SchoolNet Item Analysis Report, were important ways for district leaders to convey expectations about the use of Benchmark data. SchoolNet provided a number of online features designed to assist teachers with using the Benchmarks for instructional purposes, including multiple ways of analyzing results, links to the actual test questions, information about how to reteach the particular standards, and practice worksheets for students. To what extent these features were standardized across districts and to what extent they were tailored for the Philadelphia Core Curriculum was not clear, however. To facilitate teachers’ use of SchoolNet, the SDP issued laptop computers to teachers in the non–educational management organizations district schools over the course of 3 to 4 years. Some principals, however, 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.

Principals also accessed SchoolNet to see how students and teachers in their schools were performing for both instructional and evaluative purposes (whereas each teacher had access only to his or her class’s results, principals had access to all results within the school and could thus look for grade-wide trends). As one principal commented,

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.
Another principal reported comparing the performance of classes on the Benchmarks 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.’”

Principals also used their own review of teachers’ Benchmark Data Analysis Protocols as another opportunity to provide feedback to teachers. The principals, in turn, were required to complete Benchmark Data Analysis Protocols for their schools and review these reports with their Regional Superintendents.

Professional Learning Opportunities

The district provided several kinds of professional learning opportunities for teachers related to Benchmarks: (a) on the use of SchoolNet, (b) on the district curriculum, and (c) through 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.

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 from each school received professional development from the central office and were expected to return to their schools and train their staff. As one administrator described, “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 power point and the principals trained the staff. We wrote a script for them.” 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. In addition, the district provided professional development on different components of the Core Curriculum.

District leaders developed additional tools for teachers and school leaders as they recognized the need for other kinds of support. For example, one administrator in the Office of Curriculum and Instruction identified a need while also insisting that central office was playing a guidance 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.

Buildings also had school-based literacy and, often, mathematics 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, whereas mathematics 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.

Alongside opportunities for teacher learning, 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 ways of addressing issues of student behavior and performance. In some cases, the principals in the regions also created other opportunities to share their practices. One group of principals, for example, brought teachers and mathematics 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.

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 required time: to reteach skills and concepts, to analyze data, and to participate in professional development. The primary mechanism to support teachers’ use of the Benchmark assessments was the reteaching week scheduled after each Benchmark assessment. During this 6th week, the teachers were expected to plan and execute their reteaching, remediation, and enrichment activities. Although the Benchmark Data Analysis Protocol asked teachers how they would group or regroup students based on their analysis of the assessment results, teachers had considerable latitude around the specifics of instruction and reteaching. However, after the reteaching week ended, students moved on to the next instructional unit. The district did not provide another assessment for use during the 6th week and students were not systematically retested on the content of the instructional cycle.

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 were released around noon and teachers remained 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 to ensure that the allotted time was used to analyze and discuss student results and to learn about new instructional techniques, and this practice varied across our study schools. The principals did not have total control of this time, however, as the district directed the content of the professional development on some of the half-days.
Supports for Low-Performing Schools

The SDP implemented several additional measures to monitor and support low-performing schools that were not meeting their AYP targets. A School Assistance Team was assigned to every school that was in No Child Left Behind Act Corrective Action. A School Assistance Team was headed by a case manager who might be a retired or current district administrator and had several additional team members. The team was responsible for collecting data about a school, analyzing the data, and writing a report that was shared with the school. These data included observations in all classrooms; interviews with staff, students, and parents; and other kinds of data about student achievement, attendance, school climate, and so on. Working with its case manager, the school leadership team was responsible for developing a plan to address issues raised in the School Assistance Team’s report. The case manager and an intervention administrator worked with the school to implement the plan and monitor progress.

In 2006–07, the district created the position of School Growth Coach for the 129 schools that were in “School Improvement” status based on not meeting AYP for multiple years. 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 Coaches were 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 worked 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 predictive purposes as well, with the primary goals of assessing progress toward the learning outlined in the Core Curriculum and, ultimately, predicting success on the PSSA. 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.\footnote{In fact, the Benchmark assessment did not simulate the format of the PSSA. This quote is from an interview conducted by RFA for a separate study that predated the data collection described in the Methods section.}

The use of Benchmarks as part of the SchoolStat process and the School Assistance Team process heightened school-based leaders’ perception that Benchmarks were intended to be predictive and, indeed, part of the district’s accountability system. 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 becomes 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 tests, yet we’re treating them as high-stakes tests.

Her use of “summative” and “formative” in regards to the Benchmarks is similar to Perie and colleagues’ (2007) use of “predictive” and “instructional” purposes for interim assessments. However, some principals experienced the SchoolStat process as more focused on instructional purposes. For example, 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 [than either the results from the previous benchmark cycle or from other schools].

Discussions centered on comparing results of an individual school from one benchmark cycle to the next may miss the instructional purpose of the assessment. The benchmarks were not designed or intended to be measures of students’ progress over time. They were designed to measure students’ progress around a particular set of concepts and objectives within a single instructional period.

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. District staff raised a few issues around alignment, however. First, in a spring 2006 interview, one district administrator commented that the Benchmark assessments in mathematics were easier than the state test. Although she agreed that the Benchmarks were conceptually aligned with the state test, she argued that the greater amount of reading in the state assessment made them more challenging:

So, although we may say that it’s the same concept, it’s not the same. Because when those kids are in the Benchmark . . . 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 did not contain open-ended items, a format that was used on both the mathematics and writing portions of the state test. Philadelphia looked for a provider to supplement the Benchmarks with these kinds of items but could not find a contractor who could turn around results in a timely manner. This concern is consistent with Perie and colleagues’ (2007) discussion of some of the core characteristics of what
they argue are “effective” interim assessments designed to be used for instructional purposes; the use of non-multiple-choice items is central to their discussion.

Instructional Issues

Although 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 with support from their teacher leaders, coaches, and principals. The Core Curriculum offered extensive resources and alternative teaching strategies for helping students who experienced difficulty. SchoolNet provided information about how to reteach particular standards and additional practice worksheets for students. However, these resources were used less frequently than district leaders anticipated.

Another issue was common with state tests: The Benchmark items were given at grade level. If a student read significantly below grade level—as did many students in Philadelphia—the Benchmarks did not provide information about that student that was particularly helpful to a teacher. Beginning in 2006–07, even students whose individualized education plans required that they receive instruction off grade level were required to take the grade-level Benchmark Assessment.

Time

On paper, a full week immediately following the administration of the Benchmarks to reteach, remediate, extend, and enrich skills and concepts with students was very generous. However, before teachers could instructionally utilize the reteaching week in the ways the district intended, they had to first analyze the results of their students’ Benchmarks. In urban classrooms where 25 or more students per teacher is not unusual, this analytic task could be daunting. In practice, if a teacher administered the Benchmark on a Friday afternoon, she or he had the weekend in which to analyze the results and instructionally prepare for the reteaching week. But if the analysis and preparation did happen over the weekend, teachers had no opportunity to meet and collaborate with their grade group colleagues, an expectation of the district. Teachers’ analytic time was further compressed if the Benchmark was given on a weekday with the expectation that the reteaching days would begin the next day.

Resource Issues

As the district increasingly cut its budget due to inadequate funding, resources in support of the use of Benchmarks evaporated. For example, the bimonthly school-level professional development sessions were abolished in 2007. As school-based budgets were reduced, principals had to eliminate teacher leaders or cut back on their release time, making it more difficult for these leaders to meet with teachers about their data and suggest new instructional strategies. After the
district cancelled the SchoolStat contract, some regional superintendents tried to maintain these data meetings, but they did so without an external group to compile the necessary information.

CONCLUSION

Between 2002 and 2007, the Philadelphia school district experienced dramatic, and at times tumultuous, changes. These changes 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 assist schools and teachers. Broadly, Benchmark assessments were, from the district’s perspective, an important link between district policies and individual teachers’ classroom practices. The hope was that the assessments, along with the supports and requirements that accompanied them, would serve a primarily instructional function for teachers as they sought to improve student learning. District administrators designed the 6-week cycle of instruction in Grades 3 to 8, with the 6th week devoted to reteaching areas identified as weak through the Benchmark assessments, with the expectation that this additional time would enhance the instructional uses of the assessment.

To a lesser extent, the district also maintained both evaluative and predictive expectations for the Benchmark assessments, with higher and lower performing schools experiencing somewhat different district expectations. Assessments that fulfill an evaluative purpose by providing teachers, principals, and administrators with school-level information about how test items link to content standards can be meaningful tools for teachers. Similarly, assessments that provide educators with some predictive information on how their students might perform on a state test are also potentially useful. However, these two expectations for assessments can quickly be perceived as or become punitive if tied too directly with accountability.

In considering these multiple expectations for the Philadelphia Benchmarks, we are again mindful of Perie and colleagues’ (2007) caveat that “when an assessment system is designed to fulfill too many purposes—especially disparate purposes—it rarely fulfills any purpose well” (p. 6). It is our assessment, however, that the Benchmarks can serve instructional purposes and that the 6-week instructional cycle enhances their potential for such use. However, critical to this potential are the supports that the district put in place—easy access to user-friendly data systems, reports that focused educators on what they and their students must do to meet standards, time for analysis and collaboration, and professional learning opportunities for teachers. The recent elimination of and reduction in some of these supports, such as teacher leaders and regular district-funded professional development time, had the potential to limit the instructional benefits of the assessments. Questions around whether the Benchmarks are formative or summative assessments, and issues about how to best use the Benchmarks in altering instruction, remain critical challenges.

Other articles in this issue on Philadelphia examine these issues in greater depth. Blanc et al. and Nabors Oláh et al. look deeply into subsets of individual schools, examining how instructional communities and individual teachers have interpreted and acted on the data from interim assessments. The Philadelphia case demonstrates, among other things, that even interim assessments with a formative focus can become summative in settings in which people are differently positioned vis-à-vis power and authority and in an overall environment of high-stakes accountability and highlights some of the initiatives districts can take and challenges they face in implementing interim assessment systems.
REFERENCES


