Testing to the Test? Expectations and Supports for Interim Assessment Use

Leslie Nabors Oláh, Nancy R. Lawrence, Margaret E. Goertz,
John Weathers, Matthew Riggan, and Joy Anderson
Consortium for Policy Research in Education
Graduate School of Education
University of Pennsylvania

This paper is a work in progress. Opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the Consortium for Policy Research in Education (CPRE) or its institutional members. This document is based upon work supported by the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funder. This document has been internally reviewed by members of CPRE; however, it has not been externally reviewed.

Google "formative assessments" on a single day in July 2006 and you got 578,000 hits; by March 2007, the number was up to 783,000. Interest in such assessments is exploding, and expectations for what formative assessments can do for our students are great. The federal No Child Left Behind Act of 2001 provides the high-stakes backdrop to the growing interest in and expectations for formative assessments. In response to NCLB, districts, schools, and teachers are experimenting with interim measures intended to capture students' understanding and knowledge so that instructional action can be taken *before* more summative measures are given.

Although the rhetoric around formative assessment asserts the utility of everything from teacher-made assignments and quizzes to district-mandated benchmark testing for diagnostic and other instructional purposes, few studies have been conducted of how formative assessments are actually used. While there is acknowledgement that such assessments may be effective in improving student achievement and that students benefit from meaningful feedback, we know little about how educators use the data or about the conditions that support their ability to use the data to improve instruction.

The findings presented in this paper are drawn from an NSF-funded exploratory study of elementary school teachers' use of formative assessments in mathematics. The study also seeks to identify the ways in which policies can support teachers' uses of formative assessment for instructional improvement. In this paper we focus on expectations for formative assessment use, policy supports for assessment use; and teacher assessment practices, from the perspective of district, school, and teacher respondents in nine schools located in two school districts. Because our presentation is drawn from a preliminary analysis of our district interviews and only our first round of teacher interviews, we must stress that these findings are preliminary.

Theoretical Framework

By *formative assessments*, we mean assessments that are not primarily intended for judging outcomes and holding schools or students accordingly responsible, but instead assessments that are intended to inform instruction. We realize that formative assessments can be broadly defined 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" (Black & Wiliam, 1998). However, our primary interest is in interim assessments used for multiple classrooms. These assessments are of interest because 1) the number of schools using them is growing, as evidenced by the increase in private sector products to monitor student progress such as the Grow Network and Wireless Generation; 2) the research cited above suggests they represent promising approaches to improving instruction; and 3) as we explore in this study, we believe that policy can contribute to the effective use of such assessments. By *use* of formative assessment data, we mean how educators analyze assessment results, monitor student progress, diagnose student needs, and tailor instruction accordingly. This paper examines the uses and expectations of interim assessments by principals, coaches, teachers, and district personnel.

A number of reviews of the impact of formative assessment have been conducted, most finding substantial improvement for students in classes using some type of formative assessment. Black and Wiliam (1998) examined 250 studies of a wide range of formative assessment mechanisms and concluded that there is clear evidence that effective use of formative assessment leads to statistically significant and often substantial gains in achievement. Frequently, the gains were more substantial for low-performing students than for others, indicating that formative assessment has the potential to reduce the range of performance while simultaneously raising it on average. In a separate meta-analysis of studies using experimental designs, Kluger and DeNisi (1996) found that feedback had an average effect size of 0.4, although the range of effect size was also quite great (from $\overline{d}$ = .03 to $\overline{d}$ = .69).

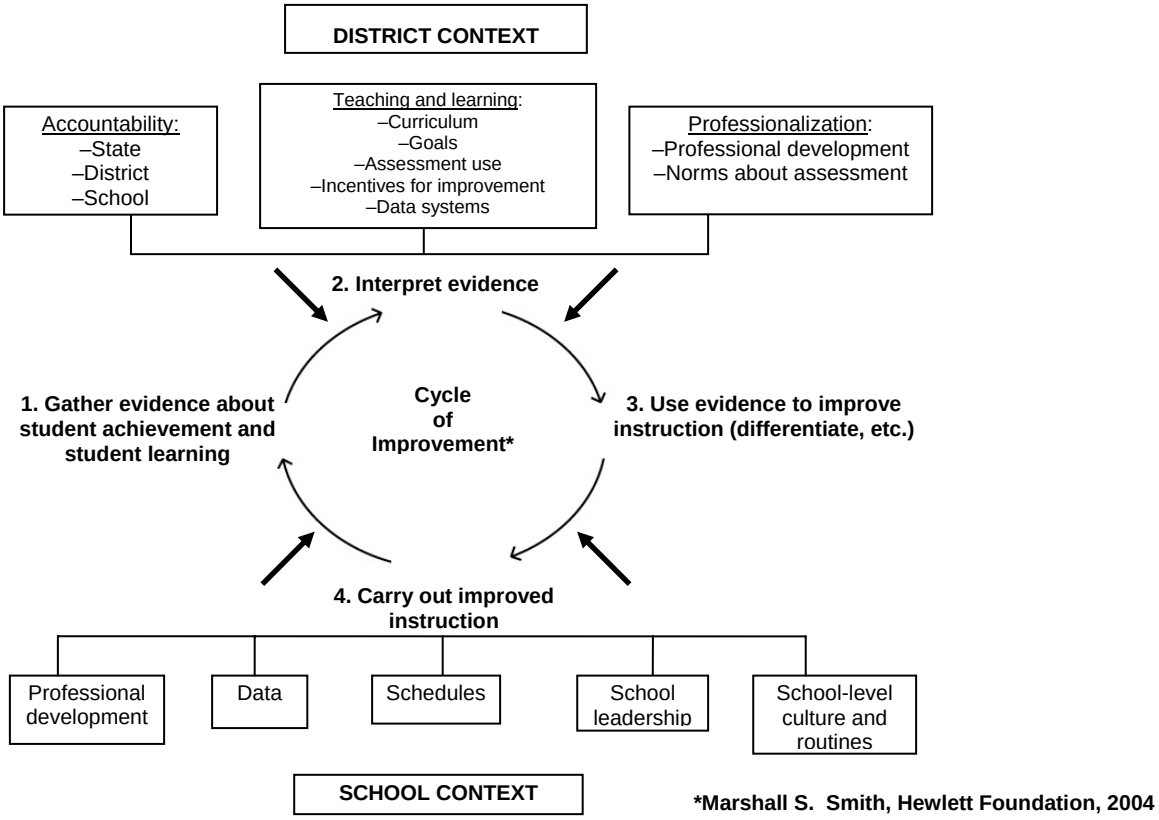
Formative assessment alone, however, is clearly not a silver bullet. Effects are highly dependent on a number of factors. Bangert-Drowns, Kulik, & Morgan (1991) found in a meta-analysis of 58 experiments that while periodic feedback generally improved student performance, the type of feedback students received had the largest effect on performance. Feedback that helped students to correct errors and reflect on the original learning goals had the greatest positive impact. Comments unique to a particular student's performance relative to an absolute standard appear to motivate students to achieve at higher levels, while responses that include solely grades or praise (or no feedback at all) seem to have little effect on student achievement, and some evidence would indicate a small negative effect from these types of feedback (Butler, 1987, 1988). In a meta-analysis of 21 studies, teachers who had distinct instructional processes to follow based on test outcomes and who had received explicit directions about how to share information with students based on the data from the assessments demonstrated significantly higher growth in student achievement than those teachers who used their own judgment about how to respond to the data (Fuchs & Fuchs, 1986).

Furthermore, recent research on formative assessment has noted several obstacles to using interim assessments formatively. First of all, it should be acknowledged that a basic tension exists between those who have the most to learn from aggregated scores on district-wide assessments (e.g., district administrators) and those who believe that looking at student work is the best way to learn about individual student competencies (e.g., many classroom teachers). While looking at student work is labor-intensive and more difficult to standardize, it has been argued that data from scored assessments tend to give only a gross sense of student performance (Shepard, 2005). In an understandable desire to limit instructional time taken for testing, districts have opted for interim assessments that are quick to administer and score. In particular, they are opting for all multiple-choice formats and for restricting the number of items given on any one assessment. Both these trends limit the use of interim assessments for formative use. Even if multiple-choice items are written to provide instructionally tractable information, such information still remains an inference made on the part of the teacher. Open-ended or constructed-response items, on the other hand, allow students to reveal their own understandings and misunderstandings. Likewise, while 20 items may be sufficient to obtain adequate reliability coefficients at the district level, the fact is that teachers want to use sub-scale scores for individual students or for groups of 2-10 students. These subscales on typical benchmark assessments can be made from as few as 2-3 items. Simply put, teacher use of benchmark assessments designed for district use can easily result in faulty conclusions born from unreliable sub-scales (Herman & Baker, 2005).

By far the most frequent obstacle that teachers see to using benchmark assessment data to modify instruction is the lack of relevant resources and support linking assessment results to teaching. For example, the third-year evaluation of Boston's FAST-R assessment system found that even though these ELA assessments provide rapid feedback on student errors, "…FAST-R is often not used to guide instruction because most of the time, it is not directly linked to curriculum and/or to the school's scope and sequence by the FAST-R coaches …"(Chrismer & DiBara, 2006, p. 4). A synopsis of recent RAND research found that while most teachers and principals reported having access to workshops on interpreting assessment results, few found them to be helpful. Educators instead preferred training on the use of assessment results in instructional planning, but this type of support "was less often available" (Marsh, Pane, & Hamilton, 2006, pp.7-8). Because of the growth of the interim assessment market and the—perhaps resultant—challenges to using such assessments to inform instruction, we believe that this topic deserves continued attention from researchers, practitioners, and policymakers.

Our theoretical framework is built on the Cycle of Instructional Improvement as conceptualized by Marshall S. Smith of the Hewlett Foundation (Figure 1). We are interested in the extent to which district authorities and schools can encourage productive use of interim assessments with various capacity-building and incentive approaches. While the purpose of this project is to explore which policy supports seem helpful, six areas seem particularly promising: professional development; data systems; schedules; incentives for improvement; leadership; and school-level routines and culture. While our preliminary findings touch on a few of these areas, we will examine all six areas in our final analysis next year.

Figure 1: The Cycle of Instructional Improvement



Methods

Sites

Our study is being conducted in two school districts, one urban and one suburban, located in one Northeastern state in the United States. The districts were selected based on a number of factors. First, to focus on formative assessments and policy supports, we held curriculum constant by choosing two districts using the same math program, in this case, *Everyday Mathematics*. Second, by studying two districts in the same state, we have held the macro-accountability context constant (i.e., in both districts the interim assessments are linked to the state standards and the state test). Third, by selecting one urban and one suburban district, we hoped to learn how policy supports for instructional improvement function in these different environments. Finally, both districts had already adopted interim assessment systems in elementary mathematics.

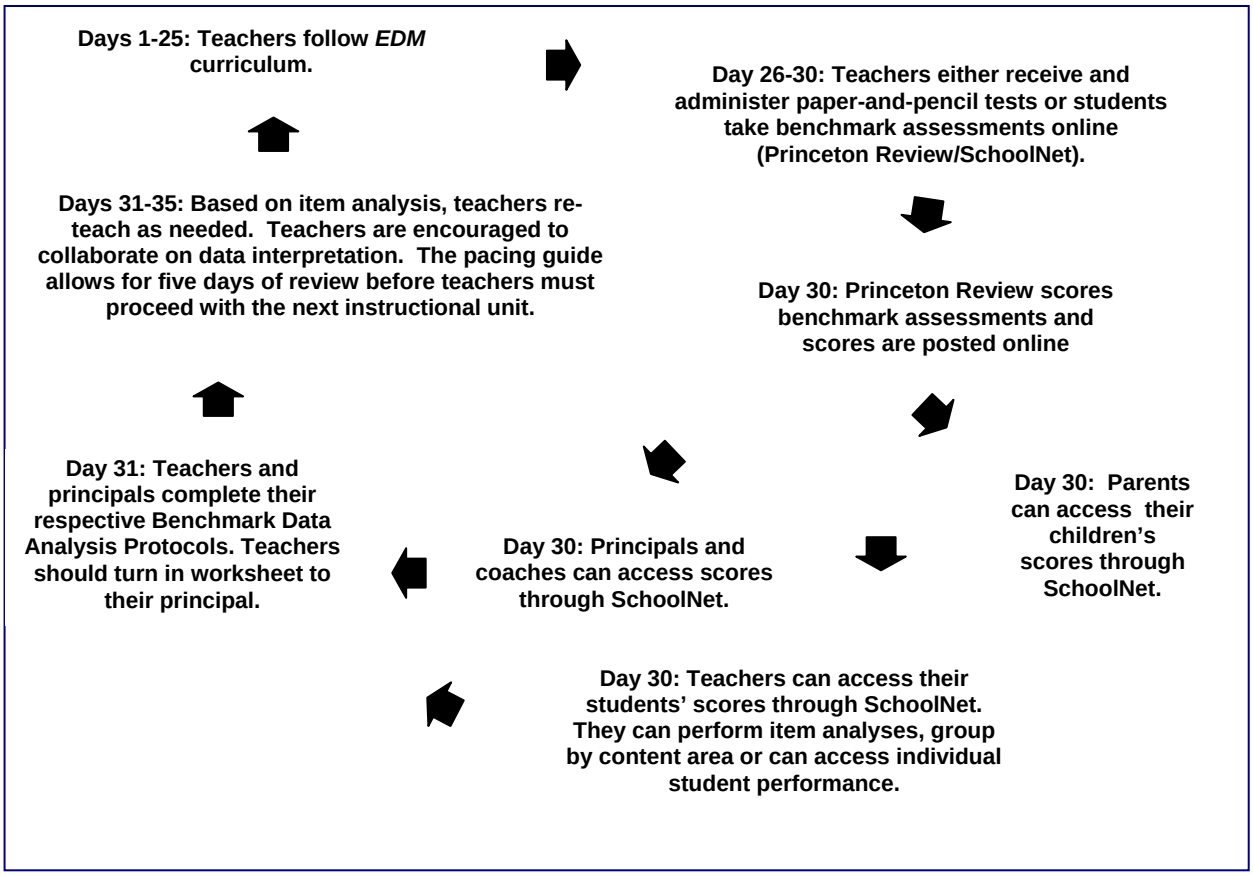
In both districts, assessments are administered to students so teachers can gauge student understanding and can take instructional action. In the urban district, a benchmark assessment is given just prior to a "re-teaching week," at the end of a 5-week instructional cycle and one week prior to the start of a new cycle. In the suburban district, a "practice test" is given approximately

three days before teachers administer a summative test, which is tied to the most recently completed curricular unit. For the purposes of this paper, we use the term "interim assessment" to include both the urban district's benchmark assessments and the suburban district's practice tests. The characteristics of interim assessments are that they "evaluate students' knowledge and skills relative to a specific set of goals, typically within a limited time frame, and … are designed to inform decisions at both the classroom and beyond the classroom level…" (Perie, Marion, & Gong, 2007). We find that these two characteristics are useful in distinguishing interim assessments from both day-to-day formative assessment practice as well as from so-called "mini-summative" assessments such as end-of-unit tests.

The urban site is among the largest school districts in the United States and has also been identified as one of the most socioeconomically, financially, and academically troubled school districts in the country. Six urban schools are included in the study; all are Title I schools. Four schools are 90-99% African-American, and the other two schools are approximately 99% Latino. In the latter two schools, the majority of students speak Spanish at home. The principals in the six schools have served their schools anywhere from 3 years to more than a decade.

Since 2003, the urban district has been using a uniform curriculum in mathematics that supports the state mathematics standards. In grades K–5, the scope and sequence of this curriculum is tightly aligned with the organization of the *Everyday Mathematics* program. The school district uses interim assessments in grades K–8 in a multiple-choice format to give teachers feedback relative to the students' mastery of the topics taught in 6-week intervals. As illustrated in Figure 2, in each 6-week cycle, the teacher is encouraged to use 25 of the 30 days for direct teaching and the other five days for review and/or extended development of topics. The 6-week period crosses units of study but is consistent with an assessment system adopted by the school district. These multiple-choice assessments are co-created by the district and Princeton Review in the weeks prior to administration and are aligned to the state's assessment anchors (and, therefore, the content of the state test) as well as to the content of *EDM*. The district has contracted with SchoolNet Instructional Management Solutions, an organization that works with districts to organize and disseminate individual and aggregate assessment data, to make assessment data immediately accessible to each teacher and family to facilitate improved instruction and communication with parents/guardians.

Figure 2: The Cycle of Instruction and Assessment in the Urban District

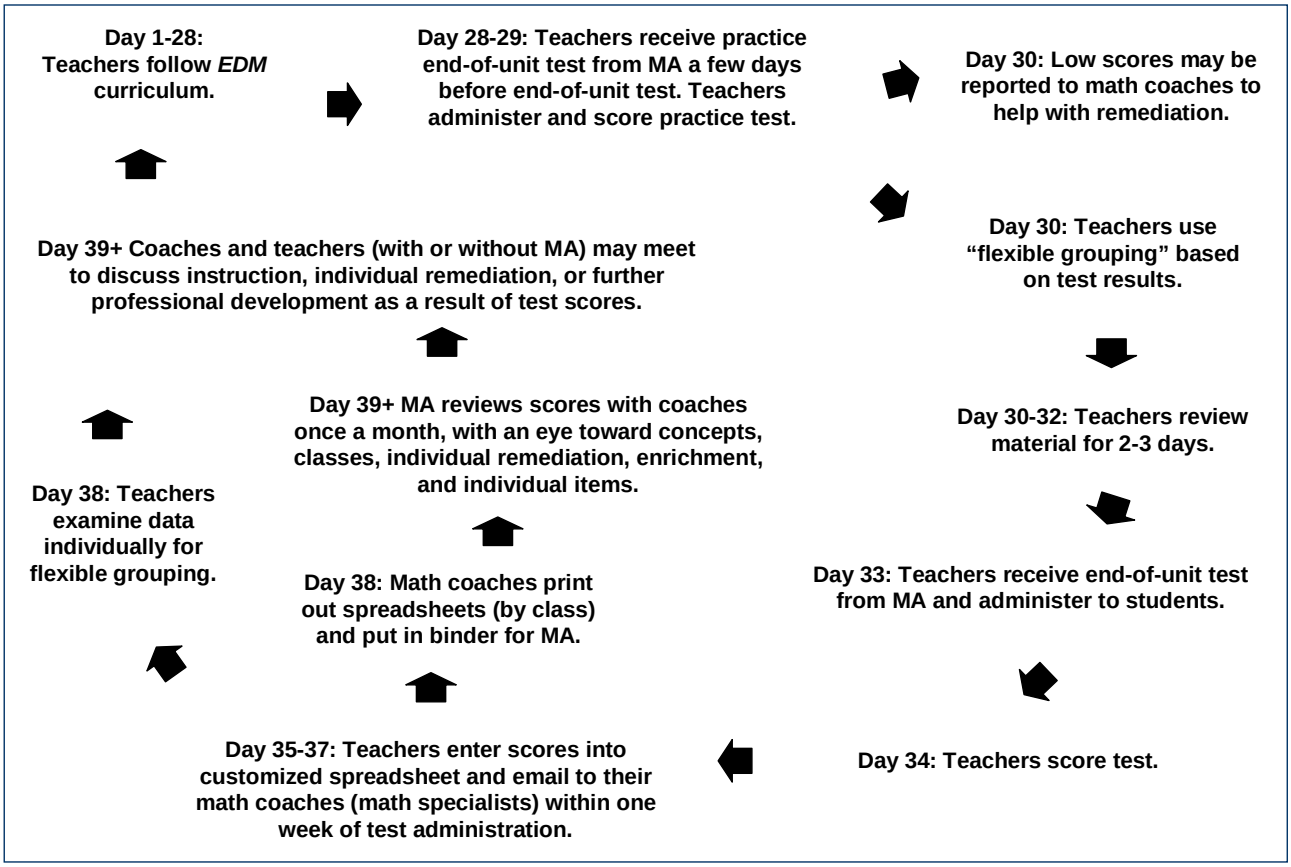


The suburban site is an economically diverse district that enrolls approximately 7,300 students in seven elementary schools, one junior, and one senior high school. Three suburban schools are included in our study, including one Title I school. All three principals have held their positions for at least 5 years, and all hold doctorate degrees.

These schools are currently engaged in a system-wide effort to implement clear and rigorous performance standards and help teachers assure that all students meet these standards. In mathematics, the district adopted the *Everyday Mathematics* curriculum in 1991 (as a pilot district), and the district leaders are strongly committed to continuing to provide standards and content-area professional development in mathematics for each specific grade at both the building level and the district level. Every 3 to 4 weeks, the district's mathematics administrator sends out to the elementary schools formative assessments that align with the pacing charts for the mathematics program. Additionally, in grades 3 through 6, the district conducts a benchmark assessment in mathematics once a year. This test reflects the state assessment anchors and models the state test. Notably, this particular test is not intended to be formative as the classroom level. Results for both assessments are collected and analyzed by school-based math specialists and discussed with the principals and teachers; in this way, math specialists function in ways similar to an elementary math coach.

The cycle of instruction and formative assessment in this district (see Figure 3) begins with approximately 28 days of instruction. The instruction is followed by the administration of the practice end-of-unit test. Teachers have some discretion on when to give the practice test but it is typically administered 1 to 3 days before the more summative end-of-unit test. Teachers receive practice end-of-unit tests from the district's mathematics administrator (MA). These end-of-unit practice tests are, in the words of the MA, "graded but not counted." Notably, students do not bring these assessments home; parents likely never see the test or might not even know it is being administered. Individual teachers are responsible for administering and scoring the practice end-of-unit tests for their class. After these practice tests are scored, teachers report the scores (by individual student) by entering them into a spreadsheet and emailing the spreadsheet to the MA. Approximately 2 to 3 days elapse between when the practice test is given and when the results of individual tests are reported to the MA. As of the 2006-07 school year, however, the district is giving principals and teachers the option of using the practice test as a pretest.

Figure 3: The cycle of instruction and assessment in the suburban district



Schools were selected according to three criteria. First, all schools made AYP in school year 2004-05. Second, although all schools met this minimum level of achievement, we chose schools to reflect a range of average mathematics performance, with 1-2 schools in each district posting district-average 3rd and 5th grade mathematics scores and 2-3 schools in each district

posting above district average 3rd and 5th grade math scores. Finally, schools were chosen to reflect the ethnic and socioeconomic diversity within each district. We focus on grades 3 and 5 as these were the only elementary grades tested by the state at the start of our study, allowing us to examine the interaction between this high-stakes, summative assessment system and the interim assessment system. These are also focal grades for elementary mathematics instruction in that it is at these levels that the mathematical performance landmarks in computation are critical for students' academic progress. Third grade typically marks the level at which students are expected to show mastery of core addition and subtraction concepts and procedures with whole numbers and of fundamental knowledge of place value. Fifth grade is the point in the curriculum when students are expected to have mastered multiplication and division and to have developed fraction concepts and skills. Fractions are crucial as foundations for continued work with rational numbers as well as algebra.

Data Sources

We conducted 24 interviews with district and school leaders, including regional superintendents; district curriculum, assessment, and technology administrators; school principals; math coaches; and school-based math teacher leaders in the spring of 2006. These taped interviews lasted approximately 1.5 hours each and followed interview protocols designed to gather information on the assessment system, the mathematics curriculum, data use, and professional development. In addition, we attended district test construction meetings, professional development sessions, principal meetings, and technology training sessions, and we have collected the actual interim assessments and other relevant documents, such as district curriculum and instruction guides.

We are observing and interviewing a total of 45 3rd and 5th grade teachers three times during 2006-07 in the nine study schools. Teacher observations and interviews are designed to capture teacher analysis of assessment results, instructional and assessment practices in mathematics, expectations for the use of formative assessments, teachers' use of formal and informal data to inform instruction, and sources of support for mathematics instruction, assessment, and data use.

All interviews are transcribed and coded in Atlas.ti to identify major themes and similarities that cut across the districts and schools. In this paper we present preliminary findings from the district and school leader interviews as well as from the teacher interviews conducted in fall 2006.

Findings

In this paper we focus on district leaders' and principals' expectations for formative assessment use; teacher assessment practices; and existing policy supports and resources for assessment use. Specifically, we address three questions:

1. What are the district leaders' and principals' expectations for interim assessment use?

2. What policy supports and resources for assessment use are available to teachers and how are they utilized?

3. How do teachers situate their district's benchmark assessments within their broader assessment practices?

Expectations

We asked district and school leaders about their expectations for the interim assessments and, more specifically, how the many different actors in the cycle of assessment (e.g., teachers, coaches, principals, etc.) are to be using these assessments and assessment results. The analysis presented here *does not* include teachers' expectations about these assessments. We are in the process of collecting these data and will incorporate teachers' perspectives in future analyses.

We found that districts and schools have numerous expectations for assessment use, and that the number of expectations for use increases as one moves from the central office into the classroom. *Districts* are expected to create instructionally actionable assessments and to use interim assessment data to identify district-wide professional development needs. Test construction seems to occupy much of the attention of district curriculum and assessment administrators; in fact, they spoke at length about psychometric properties, and both districts reported being satisfied with their current test construction processes as they have evolved.

Both urban and suburban *principals* are expected to use interim assessment results to identify children needing remediation; to identify school professional development needs; to inform their School Improvement Plans; and to collaborate with fellow principals on how to raise student achievement. Furthermore, principals are expected to follow up on teachers' data-driven instruction by, for example, observing classes.

The expectations for *teacher* use of interim assessments, however, are much more numerous. District and principal respondents offered the following (not mutually exclusive) expectations for teachers with respect to interim assessment use:

1. Identify children needing remediation and qualifying for enrichment;
2. Review the interim test with their class to provide all students with feedback;
3. Provide appropriate instruction by re-teaching to the whole group and/or forming flex groups;
4. Reflect on instruction;
5. Expand the teacher's instructional repertoire;
6. Ask school leaders for instructional support;
7. Plan or modify curriculum;
8. Communicate with parents about assessment results (if child not doing well); and/or
9. Use data as a vehicle for collaboration with fellow teachers

We also found that these expectations were fairly consistent across the two districts, with two exceptions. First, the use of assessment results to identify children qualifying for enrichment was more commonly noted in the suburban district. Urban areas may be more concerned with raising the academic performance of their lowest performing students and may be less likely to use assessment results to adjust instruction for highly proficient students. Indeed, it is an explicitly stated goal of the urban district to focus on lower performing schools as part of its reform mission. It appears that even in our restricted sample of successful schools, the district's message to focus on low performers has traction. In addition, in both districts, the interim

assessments may also be used to inform Individualized Education Plans (IEPs). The second way in which these two districts differ is in their pedagogical approaches to addressing remediation. The suburban leaders reported their teachers use flexible grouping of students to respond to differing levels of mathematics performance within classrooms, while urban leaders mentioned more varied approaches to remediation, including pull-out instruction and re-teaching to the whole class. This difference will be detailed in a later section of this paper.

Policy Supports and Resources

The major resources available to teachers in both districts were math coaches and electronic data-management systems for reporting benchmark assessment results. In addition, we found that teachers are likely to turn to their colleagues for assistance in interpreting test score data and in mathematics instruction.

Math coaches. Every elementary school in the urban district has a math coach, a former or current classroom teacher whose job, in part, consists of helping teachers make use of the benchmark assessments in math. Some of these coaches still have their own classrooms and are given limited release time to perform their coach duties; others are not grade teachers but are full-time math coaches. Until 2006, each of the 11 regions of the district also had a mathematics coach who was available to assist schools in implementing the mathematics curriculum and assessment system. Due to funding constraints, however, that position was eliminated prior to the 2006-07 school year. Every elementary school in the suburban district has a math specialist in the building. This position is officially administrative even though the math specialist may function as an instructional coach. There are a total of seven math specialists in the district. The math specialist's job is, according to the job description, "60% math," yet this position has undergone change in the past few years. Initially, these individuals were hired in part because of their mathematics content knowledge and their enthusiasm for the subject, yet now the position requires principal certification and is seen as more of a stepping stone to principalship.

Typically, in both districts, these building-level math coaches attend regional professional development sessions around math content and data analysis and are charged with returning to their respective schools to deliver training to their teachers. Thus, professional development and support for assessment use is part of the everyday work of the school math coach. These coaches, however, have many other responsibilities. District and school leaders noted the following responsibilities for math coaches in both districts:

1. Conducting school and district professional development
2. Teaching (pull-out instruction and whole-class instruction)
3. Collaborating with school psychology team
4. Discipline
5. Lunchroom duty
6. Tech support
7. Ordering materials

Furthermore, several of the urban math coaches are also classroom teachers. Coaches also mentioned that they are to do "whatever the principal wants," or that they are a "jack-of-all-

trades." These competing duties may take away from schools using their math coaches to support the instructional work of teachers. Perhaps as a result, professional development for assessment use remains undefined and focused on gathering evidence and interpreting data to the relative neglect of helping teachers to modify instruction and following up on such instruction.

Perhaps as a result of coaches' multiple duties, urban elementary teachers varied in the extent to which they see their math coach as a primary support for math assessment and instruction. For example, when asked with whom they discuss the results of their students' formative assessments or turn to for instructional or subject-matter help, for example, urban teachers were more likely to mention their grade level partners than their school's math coach. Teachers have been the recipients of coach-led professional development sessions, but teachers did not often mention the math coach as the "go to" person for support. The reasons for this are not entirely clear to us, and we will continue to investigate this trend during the remainder of our school visits.

Despite the many responsibilities of school math coaches, we did find examples of what we consider to be quality interactions between coaches and teachers around assessment results. One math coach described a conversation she had with a teacher in her school about mathematics assessment:

> *Sometimes we're able to talk it through, because I'll say, "You know, what do you think is happening here?" And a lot of times, it requires some talking it through. …I might say, "Well, how much practice did they get in this? Was it presented in homework or…in your daily math message in the same way?" You know, how much practice they have with it in the same way it was presented on the test? …And I'll say, "How many times throughout your instruction, throughout this story, did these kids have to answer, in writing, an open-ended question?" [laughter] And that's the response I get. So, and it sort of opens their eyes, because they're doing what they're told, you know, and they're doing it the way the book is telling them. And that's sometimes a problem. And when you say, you know, when they look at me and they don't understand because they did exactly what the book told them to do. And sometimes teaching requires you to go beyond that and pull from a bag of tricks.*

In this meeting, the math coach is able to guide the teacher in the importance of presenting different types of mathematics problems to students (closed and open-ended). She reminds her colleague that good teaching is more than just direct instruction (it also includes integrating homework and activities with instructional objectives). She also helps the teacher realize the potential dangers of following prescribed curriculum at the expense of using personal judgment. Perhaps most importantly, however, this coach takes an interest in the development of this teacher over the span of her career by relating the importance of acquiring her own "bag of tricks." It is also interesting to note that in this example, no scores were discussed, indicating that meaningful conversations about assessment use may be taking place outside of discussions over "item analysis" and "flexible grouping."

Suburban elementary teachers also have consistent and scheduled access to a traveling district-level math coach. This district math coach regularly visits teachers and math classrooms.

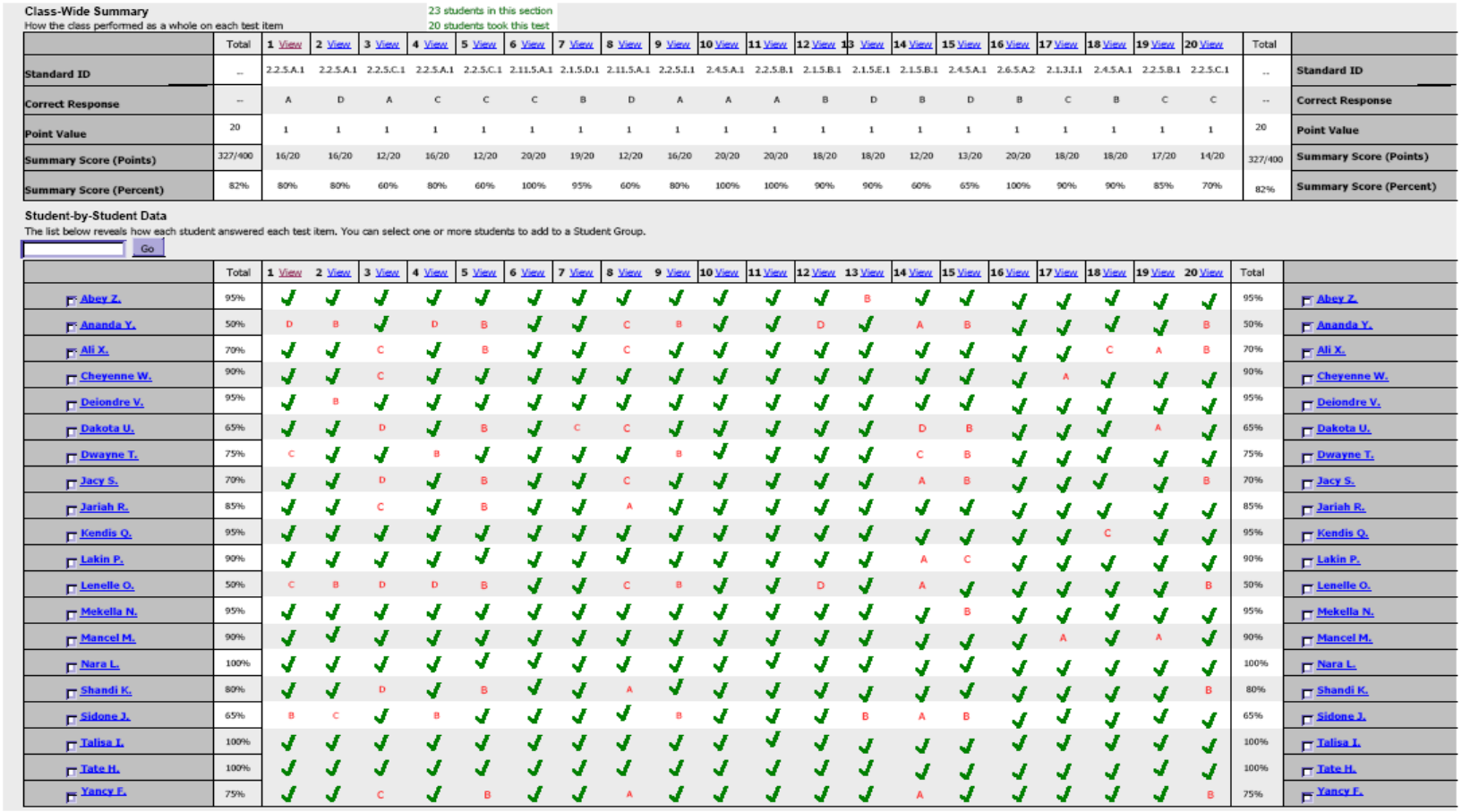
Finally, the largest of the elementary schools in the suburban district also has a dedicated math aide available to teachers.

Data management systems. One of the key features of the interim assessment systems in both districts is an online spreadsheet for teachers to record individual students' scores on the assessments. This information is accessible in both districts to the teachers and administrators, and in the urban district, to parents as well. Both spreadsheets have some minimal analysis features incorporated into the design. For the urban district, the percentage of students in each class who have missed each individual assessment item is automatically calculated along the bottom row of the uppermost panel (see Figure 4). 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 within the cell in red. The state standard to which each particular test item is linked is also noted in the spreadsheet (top row of uppermost panel). There are a host of other online features, not all used by the teachers we interviewed, but these include links to the actual test questions; information about how to re-teach the particular standard, and additional practice worksheets for students.

For the suburban district, individual test items have to be grouped together by their associated learning standard as they are entered by each teacher, and no information about the percentage of students who missed any particular item is available (see Figure 5). Teachers enter their students' practice test responses on the district-designed spreadsheet for each particular *EDM* learning goal. The district-created spreadsheet program automatically highlights cells in yellow where any student has more than one item incorrect.

Some of the suburban teachers we interviewed prefer to analyze item-level information rather than scores by content area. Consequently, we found that a few teachers in this district use much less technologically linked means of assessing areas of strength and weakness simply by recording the initials of students who missed a particular item next to that item on their master copy of the assessment. Despite district requirements to complete the online spreadsheet, at least one teacher reported that he did not follow this mandate as he felt it was unnecessary busywork.

Figure 4: Interim Assessment Results Spreadsheet for the Urban District.



Class-Wide Summary
How the class performed as a whole on each test item

23 students in this section
20 students took this test

	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	2.2.5.A.1	2.2.5.C.1	2.2.5.A.1	2.2.5.C.1	2.11.5.A.1	2.1.5.D.1	2.11.5.A.1	2.2.5.I.1	2.4.5.A.1	2.2.5.B.1	2.1.5.B.1	2.1.5.E.1	2.1.5.B.1	2.4.5.A.1	2.6.5.A.2	2.1.3.I.1	2.4.5.A.1	2.2.5.B.1	2.2.5.C.1	--	**Standard ID**
Correct Response	--	A	D	A	C	C	C	B	D	A	A	A	B	D	B	D	B	C	B	C	C	--	**Correct Response**
Point Value	20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20	**Point Value**
Summary Score (Points)	327/400	16/20	16/20	12/20	16/20	12/20	20/20	19/20	12/20	16/20	20/20	20/20	18/20	18/20	12/20	13/20	20/20	18/20	18/20	17/20	14/20	327/400	**Summary Score (Points)**
Summary Score (Percent)	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.

Go

	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	
Abey Z.	95%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	B	✓	✓	✓	✓	✓	✓	✓	95%	Abey Z.
Ananda Y.	50%	D	B	✓	D	B	✓	✓	C	B	✓	✓	D	✓	A	B	✓	✓	✓	✓	B	50%	Ananda Y.
Ali X.	70%	✓	✓	C	✓	B	✓	✓	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	C	A	B	70%	Ali X.
Cheyenne W.	90%	✓	✓	C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	A	✓	✓	✓	90%	Cheyenne W.
Deiondre V.	95%	✓	B	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	95%	Deiondre 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	✓	A	✓	✓	✓	✓	✓	B	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	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	B	80%	Shandi K.
Sidone J.	65%	B	C	✓	B	✓	✓	✓	✓	B	✓	✓	✓	B	A	B	✓	✓	✓	✓	✓	65%	Sidone J.
Talisa I.	100%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100%	Talisa I.
Tate H.	100%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100%	Tate H.
Yancy F.	75%	✓	✓	C	✓	B	✓	✓	A	✓	✓	✓	✓	✓	A	✓	✓	✓	✓	✓	B	75%	Yancy F.

Figure 5: Interim Assessment Results Spreadsheet for the Suburban District.

Practice test problem numbers	19,20,21,22	3,4,5,6,7,8	9,10	11, 12, 14	1,2,13	15,16,17,18
Learning Goal	Equivalent mixed numbers	Adding & subtracting fractions and mixed number	Percent-decimal-fraction correspondence	Comparing or ordering fractions	Finding common denominators	Multiplying fractions
Proficiency level	S	D/S	S	D/S	S	D
Number of items	4	6	2	6	2	4
Name	Number wrong on Practice test					
1. Michael Ambruster		1		1		
2. David Bridgewater	1	2	1	4	2	1
3. Brittany Cooper				3	1	
4. Skye Davidson		1				
5. Hodgkin Eames	1		1	1		
6. Paige Fairly	1	5	2	4	2	1
7. Tony Garafalo				1		
8. Sorrell Hill	3	2		2	1	
9. Madelaine Isaak	4				1	
10. Alexander Jacob				1		
11. Kiki King			1	2	1	
12. Anton Lang	2	4	2	3	2	1
13. KC Monroe						
14. Clay Nailor			1			
15. Daniel Ooster	1	1	1	2	1	1
16. Adam Powell						1
17. Elif Ross						
18. Jenna Smith	1		1	2	1	1
19. Randal Tatum		1			1	
20. Ari Urbinski			1	3	1	
21. Jonah Valdez	4			2		
22. Ambrosia Wallace		1	1			
23. Ynes Yaragosa				1	1	

While teachers mentioned benefits of the data management systems in both districts (centralization of information and ease of data retrieval being two commonly mentioned ones), there has been growing tension in the urban district about online test administration. In the past two years, students have taken the test on paper and then typically, the teacher received some time at a computer lab for the students to then 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. The curriculum department, in turn, has fought to maintain the option of having students take the assessments in pencil-and-paper format as they felt that the computer format was poorly executed and required extensive scrolling up and down to see all of the components of each question. Because of this fine-motor task demand, elementary school students, it was believed, may not accurately complete questions even if they possessed the appropriate skills or understood the requisite concepts. As of this writing, this issue has still not been completely resolved, but for the 2006-07 school year 3rd graders have received a dispensation from this online test administration policy. Because of the perception that students may make careless errors while entering their scores, a few teachers reported that they enter the answers into the computer off of student answer sheets. These teachers mentioned that they don't want students to make any "mistakes."

Regardless of the time commitment involved, teachers mostly believe that the online spreadsheet systems provide information on student performance beyond what their own assessments and analysis would provide them. The core of teacher comments regarding this value spoke to how the spreadsheets make it clear exactly who and how many missed a particular question or a particular standard. Below, an urban teacher is asked whether she gets important student information from the benchmarks that she does not get from other sources:

> *Well, I definitely like the breakdown how you can see what the whole class did on a specific item, instead of just having to go through your tests on your own. You can see the percentage that the class got, and that's a big help.*

It is interesting to note that most of our teachers see the main contribution of the data management system as organizing and clarifying interim assessment results, but not necessarily assisting teachers in interpretation of these results or planning for re-teaching. In the urban district, this trend exists in spite of the fact that the SchoolNet Align™ feature also offers curricular links and an instructional planning guide. Still, these teachers see the data management tool as offering useful information that they might not otherwise be able to easily access. Below, we include an exchange between a researcher and another classroom teacher in which the researcher is probing on how the teacher uses SchoolNet Align™:

> Researcher: *Would you have known you wanted to review those areas with the—Or did it provide a clearer sense of what you—*
>
> Teacher: *It most definitely was clearer, because it tells you the percentage right there. And you can even zero it in even more. And it'll tell you exactly what students missed that question. It's just good.*
>
> Researcher: *And you couldn't have created that same information yourself, necessarily?*

> Teacher: *No, because when I give, for example, the unit test, it has many different skills on the test that you're testing all at one time. So, I can look at it and we talk about when we—how we grade it and we block each section, but you're looking at it overall, not specifically student by student. So it does make it clearer.*

Other sources of support. Teachers' grade group colleagues were the most often cited source of collaboration and support for teachers in both the urban and suburban districts. Teachers were quick to turn to their grade-level colleagues to discuss the results of the interim assessments, to share strategies, and to seek advice. According to a 3rd grade suburban teacher:

> *We sit and talk about where we are and what we see and what we would like to change or what we'd like to add. Often times, we bounce ideas off each other. And if we have a particular group that is struggling, sometimes one teacher's idea is the light bulb for another.*

However, in spite of the value teachers placed on talking with colleagues, it is not dedicated, structured time. When a 5th grade suburban teacher was asked if the grade-level conversations she described were formal ones, she said, "Oh, not at all. We're not given collaboration time like that. We'll just get together over lunch and talk … ." Shared lunch periods provide many teachers with unscheduled, but highly valued, opportunities for teachers to discuss the interim assessments with each other.

A few urban teachers in our study also utilize student teachers and volunteers as support for classroom instruction. Such support can expand the re-teaching options that a teacher is able to pursue. Below, for example, a 5th grade teacher describes how she is able to meet with her students needing additional instructional support because she has a student teacher in the room:

> *I try to spend a lot of time working with kids individually. And having a student teacher, it's great because she's here Tuesdays and Wednesdays and so we do more working in pairs or having them work at their seats or in groups and then the two of us walk around and we're checking how they're doing, and re-explaining the concepts. And so this way, I can see every student on an individual basis and I can have them explain to me certain things, "How did you come up with this number? What was your thinking?"*

With respect to volunteers in the classroom, the urban district has a program that brings retired individuals into the classroom. It is not clear to us how these volunteers are assigned to specific classrooms, how many volunteers there are, or what their formal classroom duties are. The small handful of teachers with these volunteers in the classroom have used the volunteers to sit with a small group of students identified as needing additional instructional help at the back of the room during instruction. Because this district currently has a shortage of teachers and is in the middle of a teacher-recruiting drive, some classrooms are assigned student teachers for a semester at a time. However, few teachers have student teachers or volunteers available to them, so their ability to work with small groups varies considerably within and across the urban schools we visited. It should be noted that while student teachers and volunteers can ease a teacher's workload and reduce the student:adult ratio in the classroom, by definition, it is highly unlikely

that they are either certified to teach or have substantial classroom experience. Therefore, while the suburban teachers are able to bring math coaches and math aides into their classrooms, the urban teachers are more likely to rely on less-experienced helpers.

Teachers' Use of Benchmark Assessments

While we are in the process of collecting information on teachers' expectations for the use of benchmark assessments, we did look to our first round of interviews to get a sense of how teachers situate their district's benchmark assessments within their broader assessment practices. We found that benchmark tests are one of many ways that teachers assess their students' knowledge of mathematics, in addition to their day-to-day assessment of student understanding, weekly tests, and end-of-unit *EDM* tests. While common wisdom suggests that teachers are overwhelmed by the testing regime and, as a result, loathe the addition of new assessments, most of the teachers in our sites found interim assessments to be a valuable tool for gauging student performance in mathematics.

As a general rule, the primary function of each district's assessment tool is diagnostic—ostensibly providing data to teachers about how well or poorly their students mastered the material covered in an individual unit or instructional cycle. While in many cases teachers reported having a good sense of how their students would perform prior to administering the assessments (especially in the suburban district), teachers still rely heavily on the aforementioned assessment data to inform their remediation/re-teaching time and overall assessment of student progress:

> *We are mandated to give those* [practice tests] *a few days in advance, determine where the areas of weakness are, and then work on those areas of weakness prior to the test in order to support those children in an effort for them to succeed in a better way on their test…. I use it faithfully. And it really is a good tool for that.*
>
> *Well, for example, with the benchmark assessment, it clearly delineates what skills are lacking. Like, it will show you … if a majority of the class got a particular question wrong.*
>
> *The day they take it [the practice test], I take it home and do it that night so that the following day is when we can meet, reinforce certain things, re-teach certain things.*
>
> *I can print out an item analysis, and I can see, OK, out of 23 kids, 18 of them got the question wrong that asks them about elapsed time. Well, OK, on our review week, guess what? We're going to be doing elapsed time.*

The benchmark and practice tests prove useful in other ways. For example, they help teachers keep track of the content they are expected to cover and to reflect on the degree to which they are doing so. As one teacher observed, "The benchmark assessments have spoken more to me about what it is I'm teaching … rather than which students are getting it or not." Another teacher suggested that the benchmark assessment "sort of keeps us in tune with what [the students] are supposed to know."

Another benefit of the benchmarks and practice tests was that they confirm, and in some cases formalize, information generated by teachers' informal assessment practices. Some teachers suggested that the interim assessments do not necessarily tell them anything new about their students, noting that their more frequent assessment processes lead them to conclusions that mirrored those of the benchmark results. "I can't say [the benchmark assessment results are] a big surprise," one teacher commented, "because as we're going through the *Everyday Math* we kind of know where kids are, if the interest is there, if the hands are up. You kind of know if you've got them if they're understanding it." Another teacher remarked, "I know in my head who is struggling and who is not." The contribution of practice tests, however, is that "it's in writing now that we know who's struggling rather than before it might not have been recorded formally." The interim assessments, therefore, might serve an important validating function even when if they do not add to teachers' understanding of student strengths and weaknesses.

Teachers use the interim assessments not only to identify areas of student weakness but also to determine what type of performance constitutes weakness. As mentioned above, the suburban district's data management program automatically identifies students falling above or below a certain threshold, enabling teachers to identify students in need of additional help. As one teacher in this district explained,

> *After we give the pretest, we actually have a computer-generated type of program where we put in the number students got correct under each area. It will come up yellow. So it's there. It's yellow. It tells you, OK, so-and-so didn't understand this concept....*
>
> *So, it'll say "Numbers and Computation" were problems 1 through 6. And I put in how many they get correct. And if it's below, it automatically highlights in yellow. It takes a lot of work off my back, because I used to have to do that, decide, well, what's below, what needs help?*

Teachers in the urban district also discussed using proficiency cut-points to identify students in need of assistance. One school, for example, categorizes students as "proficient," "strategic" or "at risk."

> *We code them green, yellow, pink—green being target, or mastery; yellow, strategic; pink, at-risk below level…. The 90 to 100 for their percent would be a mastery or proficient.…The 70 to 85 would be where I would consider strategic, like the middle of the road. And the 65 and below definitely would be the children that would need some type of small-group instruction on what some of their weaknesses are.*

Other teachers appear to set their own thresholds for defining when re-teaching or remediation is necessary, such as students who got fewer than 70% to 80% of the items correct.

Ironically, although most of our teachers reported using the results from the interim assessments to identify weak content areas or lower performing students, recent changes in the urban district's test administration procedure has rendered those interim assessments virtually

useless for teachers of special-needs children. Prior to the 2006-07 school year, students in special education classrooms took the interim assessments at instructional level while taking the state test at grade level. This arrangement seemed to make sense as the results of the interim assessment were intended to inform classroom instruction and learning, while the state test was to track all students' progress toward a grade-level standard. In 2006, however, the urban district changed this policy, instead requiring all children with IEPs to take the interim assessments at grade level. As a result, special education teachers in the urban district now find the interim assessment results meaningless, as all of their students score at the floor of the assessment. In the words of one special education teacher:

> *Last year they gave them all the 3rd grade benchmark. And it was wonderful. They could do the work. This year it's on grade level, and it's so frustrating because there's four grade levels in here. So, it takes a lot longer and I do a lot of guided stuff, because they look at the problems and they have no idea what the problem is even asking for. I have a 6th grader functioning on a 4th grade math level trying to take the 6th grade benchmark and he can't….They've never done average before, and there was an average question. So, to tell them how to do the average to get it and they could do it. But the results don't help me any at all.*

In the suburban district, where the interim assessment is given to special-needs students at grade level, all special education teachers who were interviewed reported using the results to identify content areas that periodically need review. This is particularly interesting in that special education teachers generally have a larger repertoire of assessments at their disposal and in that students with IEPs are generally subject to more frequent assessment using various types of instruments. In spite of the case that periodic, diagnostic assessment occurs more often in special education classrooms, these teachers still find value in the district's interim assessments when they are given at instructional level.

We found that while the use of interim assessments to identify content or skill-related weaknesses was similar across the two districts, the districts differed in how teachers responded to low-performing students or to weak content areas. In the suburban district, teachers echoed administrators' commitment to flexible grouping as the primary instructional grouping technique for addressing student's different learning needs. Most of the suburban teachers reported using flex grouping in response to both interim assessment results as well as in response to their other short-cycle formative assessment practices. In this way, individual student performance on the practice test on particular content areas is used to form the groupings. This was seen as an efficient way to address student learning needs. As one teacher explained:

> *…So with the flex group, it's definitely more targeted. So you're working on a specific, focus-targeted skill. So even in your flex groups, you're not taking--"OK. I'm going to take the bottom three kids, and then we're just going to sit and we're going to work through the whole test." That's not the purpose of it, because even those bottom three, or whatever, there's something that they did well on that they don't necessarily need review in.*

Some of these teachers have even created new "kid watching" charts or "flex grouping sheets" that help them keep track of their flexible grouping from day to day. It is important to note that while the suburban teachers mentioned flex grouping the most frequently, they see it as one of several instructional options for re-teaching material. In this way, flex grouping is treated as a "first among equals" among other instructional approaches in the suburban district.

Teachers in the urban district also reported using different strategies to address student weaknesses, but the most common approaches were to re-teach to the whole group using the items on the benchmark assessment (or similar items) as a basis or to provide individual attention either by assigning extra work or by pulling children out for remediation. The difference in approaches is usually determined by the number of students who get any one item incorrect. As one teacher explained, "… depending on the skill … I would do a small-group instruction, or maybe I'll send home specific homework for each child that needs—if it's half the class, I might as well—I'll just re-teach the whole thing. But if it's a few children like this, then I would definitely pull them out and get some special homework for them, for them to work on."

In any case, most teachers in the urban district reported that they would consult the state standards prior to re-teaching. When one teacher discussed her benchmark assessment results, she mentioned, "I'm surprised more of them didn't get the symmetry wrong, because that's hard…. I would just make a note and go back to the standards and just write everything down. And then I would re-teach and reassess." Teachers also reported accessing a variety of additional materials to provide students with practice in skills on which they were seen to be weak. One teacher described this process of linking the assessment information to resources as such:

> *I would look at the different standards that the kids, as a group, did not understand, and I would re-teach those for standards. I would look them up and make sure I saw what they were, and then maybe either look in the teacher's manual for more ideas, because sometimes there's extra activities we didn't get to, or I could look on the Internet for different ones, for more practice.*

Conclusion

Perhaps the most important finding to emerge from our preliminary analysis of interim assessment use is that such assessments *are being used* by teachers, principals, and district leaders. This seemingly simple finding is not insignificant. Our preliminary findings indicate that teachers' and districts' use of interim assessments has moved beyond merely administering and scoring the assessments to attempts at incorporating the knowledge gained from these assessments into teachers' assessment practices, including instruction. What we have learned is that, to varying degrees and depths, teachers and district leaders analyze, discuss, review, reference, and consider the mathematics interim assessments with an eye toward re-teaching and reviewing their students' weaker concepts and skills.

To support such activities, both districts have restructured their school calendars to allow for re-teaching and reviewing of mathematics. In the urban district, a full week is dedicated to the re-teaching of mathematical concepts. In the suburban district, teachers are given 1 to 3 days in which to review mathematical ideas before a summative test is administered to students. In

rearranging the schools' schedules, the two districts have made a considerable commitment to interim assessments. However, this is not to suggest that more cannot be done in both districts to further support teachers' use of interim assessments. We have also found that most teachers lack dedicated time during the school day to meet with their grade-level colleagues to discuss and analyze the interim assessments. Currently, it appears that such grade-level conversations are irregular and not institutionally supported. Furthermore, vertical conversations, that is, conversations about the interim assessments across grades, are rare. The spiraling curriculum that characterizes *Everyday Mathematics*, and is used in both districts, suggests that teachers would benefit from having such conversations with teachers teaching in grades below and grades above them. Our preliminary findings also suggest a need for increased professional development around elementary mathematics content and data analysis. Without additional professional development, the districts' days devoted to re-teaching and review are vulnerable to instructional repetition. By adding to a teacher's instructional repertoire, the likelihood of her successfully reaching students at varying stages of learning increases.

Remarkably, given the current high-stakes testing climate that has come to define many of our public schools, we encountered little or no administrator and/or teacher resistance to the interim assessments. Instead, teachers appear to look to the interim assessments to corroborate what their other informal assessment practices tell them about their students. Put another way, teachers seem to trust what the interim assessments reveal. And in trusting the measurements, the interim assessments occupy an important location in the cycle of instructional improvement. The one notable exception to this trend is the special education teachers in the urban district who have seen instructional-level interim assessments replaced by grade-level counterparts. This move has rendered these assessment results meaningless for classroom-level use.

While we stand by the preliminary findings of this study, we are cautious not to paint too rosy a picture of interim assessment use in the two districts. The districts have been successful in the first phase of the instructional improvement cycle in that they are gathering evidence about standard achievement and student learning. In addition, the two districts have taken important steps in the second phase—that of interpreting evidence. However, the more challenging work, and the work that holds the most potential for student learning, occurs in the last two phases of the cycle: using evidence to improve instruction and implementing improved instruction. It is here, in these last two phases of the cycle, that both districts will need to dedicate both time and resources.

References

Bangert-Drowns, R. L., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213–238.

Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment.* London: Kings College.

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology, 79,* 474-482.

Butler, R. (1988). Enhancing and undermining intrinsic motivation; the effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology, 58,* 1–14.

Chrismer, S.S., & DiBara, J. (2006). Formative assessment of student thinking in reading (FAST-R). Cambridge, MA: Education Matters, Inc.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53,* 199-208.

Herman, J. L. & Baker, E. L. (2005). Making benchmark testing work. *Educational Leadership, 62,* (3), 48-54.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback intervention on performance: A historical review, meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254-284.

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). Making sense of data-driven decision making in education. Santa Monica, CA: RAND Corporation.

Perie, M., Marion, S., & Gong, B. (2007). A framework for considering interim assessments. Unpublished manuscript.

Shepard, L. (2005). Formative assessment: Caveat emptor. Paper presented at the ETS Invitational Conference, New York, NY.