DRAFT REPORT- Please do not cite or distribute without permission of authors

Learning to learn from benchmark assessment data: How teachers analyze results

Leslie Nabors Oláh, Nancy Lawrence, and Matthew Riggan

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

Annual Meeting of the American Educational Research Association
New York, New York
March 27, 2008

Research for this paper was supported by a National Science Foundation Grant (#REC-0529485) to the Consortium for Policy Research in Education at the University of Pennsylvania

This research was made possible, in part, by the support of the School District of Philadelphia. Opinions contained in this report reflect those of the authors and do not necessarily reflect those of the School District of Philadelphia.

I. Introduction

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 acknowledgment 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.

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. From an efficiency standpoint, this makes sense. The question is how these interim assessments are used by teachers.

The analysis presented here is part of a broader research agenda developed by the Consortium for Policy Research in Education (CPRE) to better understand how teachers, schools, and policy makers can use information about student learning to inform decision making and practice. CPRE houses the Center on Continuous Instructional Improvement (CCII)[1], a center that provides leadership in research and development to improve the quality and expand the use of policies, systems, and tools that support three closely related improvements in public education: adaptive instruction, formative assessment, and the cycle of instructional improvement. The findings presented in this paper are drawn from an NSF-funded exploratory study of elementary school teachers' use of interim assessments in mathematics. We use the term "interim assessments" to refer to assessments that (a) evaluate student knowledge and skills, typically within a limited time frame; and (b) the results of which can be easily aggregated and analyzed across classrooms, schools, or even districts (Perie, Marion, & Gong, 2007). As mentioned above, this type of assessment is becoming increasingly popular as a way of informing teachers, schools, and districts about student performance. This paper addresses the question: *How do the Philadelphia teachers in our sample analyze benchmark assessment results, how do they plan instruction based on these results; and what are their reported instructional responses to such results?*

II. Site Description and Assessment

Philadelphia is among the largest school districts in the United States with over 214,000 students (K-12) enrolled in 276 schools. It has also been identified as one of the most socioeconomically, financially, and academically troubled school districts in the country. Since 2003, the School District of Philadelphia has been using a "Core Curriculum" in mathematics that supports the Pennsylvania 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 Core Curriculum follows a tightly sequenced "Planning and Scheduling Timeline," or pacing guide, that details the content and types of activities that each lesson should include. With

[1] The Center on Continuous Instructional Improvement is funded by the William and Flora Hewlett Foundation.

the adoption of the Core Curriculum, the district also instituted a benchmark assessment system designed to give teachers and principals feedback on student performance every six weeks. These multiple-choice assessments are cocreated by the district and Princeton Review in the weeks prior to administration and are aligned to the Pennsylvania Assessment Anchors (and, therefore, the content of the PSSA) as well as to the content of *EDM*. In 2004 the office of Curriculum and Instruction created two-page "Benchmark Data Analysis Protocol" worksheets, one version for teachers and one version for principals. These were designed to help teachers and principals in their analyses and planning following each benchmark test. The teacher version focused on identifying students' weak areas, regrouping students, changes in teaching strategies, and subsequent testing for mastery. The principal version emphasized the school's weak performance areas, discussion of data, and further action steps.

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 six-week intervals. In each six-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 six-week period crosses units of study but is consistent with an assessment system adopted by the school district. The 20-item, 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, and to make assessment data immediately accessible to each teacher and family to facilitate improved instruction and communication with parents/guardians.

III. Sampling and Data Sources

Schools were purposefully selected according to three criteria. First, all schools made AYP in school year 2004-05, the planning year of the study. 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 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 the School District of Philadelphia. We focused on grades 3 and 5 as these were the only elementary grades tested by the state at the start of our study. 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.

For the larger study, we collected data from six sources: classroom observations; teacher interviews; school and district leader interviews; observation of district and school meetings; artifacts; and a teacher survey: Content Knowledge for Teaching – Math (CKT-M). A

description of the district context is presented in our colleagues' paper, *Building With Benchmarks: The Role of the District in Philadelphia's Benchmark Assessment System*. The preliminary findings discussed in this paper are based on three rounds of teacher interviews (conducted in Fall, Winter, and Spring), observation of district and school meetings, and relevant artifacts. These three sources are described in detail below.

Teacher Interviews

In the Fall and Winter visits to schools, we conducted individual, hour-long interviews with teachers immediately following each classroom observation. In almost all cases, these interviews took place right after, or a couple of hours after, the observed lesson. Our Spring interviews with Philadelphia teachers, however, took place two weeks after classroom observations due to the administration of the state test (the PSSA) in the days immediately following the classroom observations. All teacher interviews were audio recorded and transcribed.

Fall interviews. The Fall teacher interviews consisted of two parts: semistructured questions and a Data Analysis Scenario. The questions focused on teachers' professional backgrounds, their general assessment practices, and the professional development opportunities available to them. We also asked several questions that helped provide context for the lesson that we had just observed and that were designed to tap into the different ways in which teachers monitor student understanding of mathematical content. We also asked the teachers if there was anything that they struggled with "mathematically" during the lesson.

The Data Analysis Scenario consisted of a hypothetical mockup of student results based on each grade's interim assessment. The items on each of these two Scenario versions were taken directly from the district's original interim assessments following a unit on fractions. We presented teachers with these hypothetical interim assessment results for two reasons. First, at the beginning of our study, we did not know the extent to which participating teachers used the district's assessments results, the district's reporting mechanisms, or both. At this early stage in our relationship with the teachers, we felt it would be too intrusive to ask them if they would be able to discuss their own students' results with us. Therefore, in order to learn more about teachers' familiarity with their district's assessment system, we presented them with a basic report (the Item Analysis) from a hypothetical class of students. The second advantage to using a hypothetical set of results was that we could standardize the "results" across grades (and districts as part of the larger study) to see what variation in teacher analysis or interpretation would occur in response to an identical set of results. For example, while there might be some teachers in our sample who rarely see incorrect answers on their students' actual interim assessments, we wanted to see how *all* of our participating teachers *would* respond to certain patterns of incorrect responses. Therefore, we designed each of the Scenarios so that 82% of the items are correct, and incorrect responses reflect common student misconceptions in mathematics (e.g., the incorrect responses to the Philadelphia interim assessment indicate that several students cannot find common denominators, leading to errors in comparing and adding fractions).

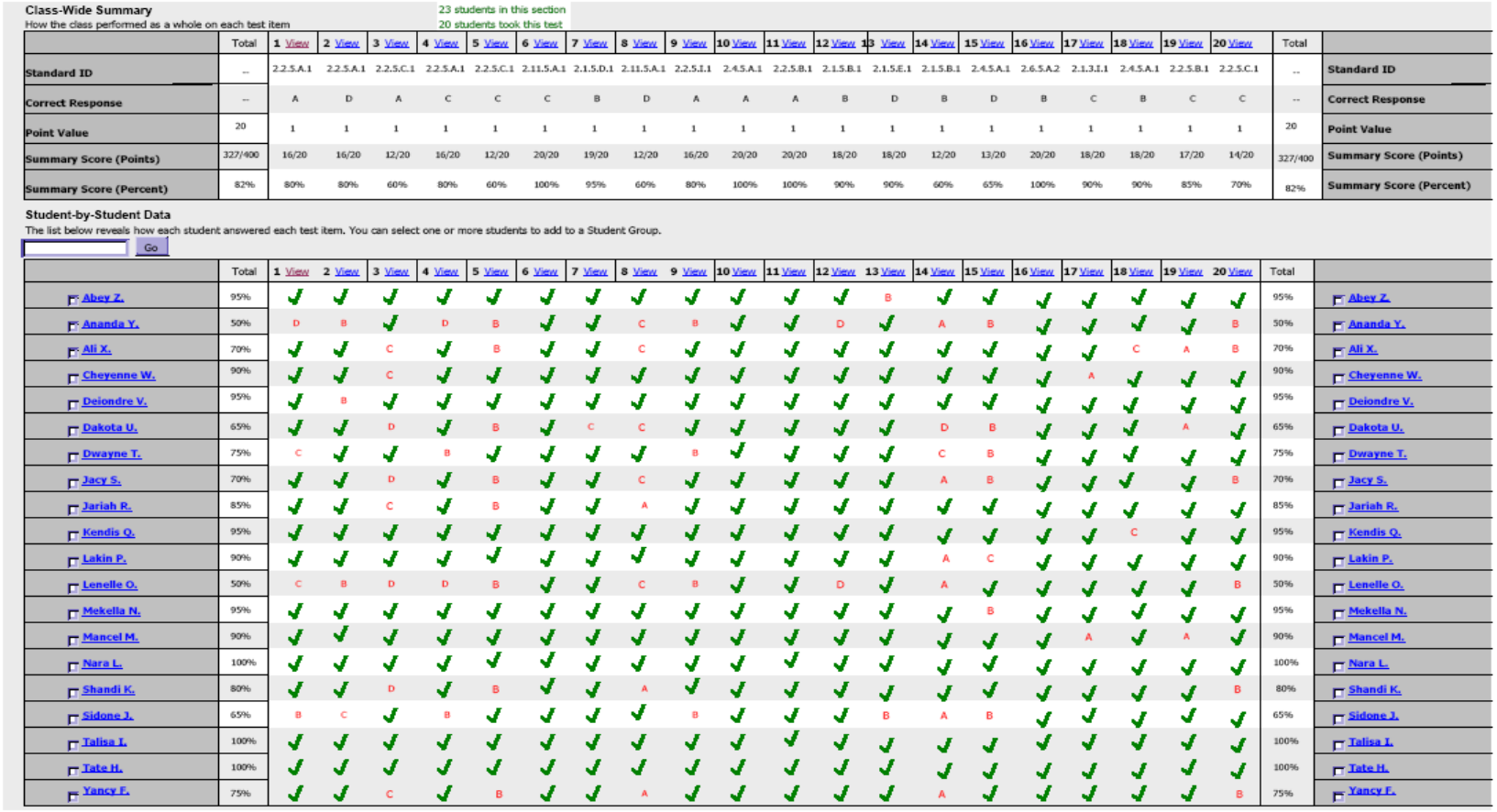
The Data Scenarios were formatted so that they mirrored how the School District of Philadelphia reports their interim assessment results. In Philadelphia, teachers receive their

students' scores online through SchoolNet. Since we did not want the Scenario to become unwieldy to administer, we chose to present the Philadelphia teachers with a printout of only the most commonly accessed view of student results, the Item Analysis (see Figure 1). Since the Philadelphia online data is stored in a Microsoft Access database, we used the same database management software to create the Philadelphia Data Scenario. The Item Analysis view is constructed much like a grade book, with students' names listed down the left-hand side. Each column represents individual student performance on each of the 20 items; green checkmarks indicate correct answers, while red letters indicate which incorrect answers were chosen from among four answer choices (i.e., A,B,C, and D). Percentages correct are reported both by student and by item; for example, from looking across the Summary Score row we can see what percent of the students in the class scored correct on each item. The percentages shown next to each student's name, on the other hand, report the proportion of items that the student succeeded in answering correctly. Two more features are worth noting: First, in the top panel, the Standard ID row shows which Pennsylvania state standard each item is aligned to. Second, while we presented teachers with a color "printout" of these hypothetical results, were teachers to access their class scores online, certain features of the spreadsheet would allow them to "drill down," or link to other information. For example, clicking on the View link above an item would take a teacher to a screen displaying the actual benchmark assessment problem.

During the Fall interviews, we presented each teacher with a one-page printout of hypothetical interim assessment results, asking the teacher if she "had seen something like this before." In all cases, teachers reported having seen their district's interim assessment results reported in this way. We then asked each teacher to imagine that this was her class and to "think aloud" for us about what she saw in the results. After approximately five minutes, or after the teacher stopped talking, we continued with a series of six follow-up questions designed to call attention to patterns in the data (e.g., *Are there any topics that this class, overall, appears to have difficulty with? How do you know?*). In this way, we were able to capture both each teacher's initial, natural reaction to the assessment results as well as whether or not, with probing, she noticed particular strengths and weaknesses among "her" class. Finally, we presented each teacher with two "misconception probes" based on actual items from each interim assessment. In each case, a hypothetical student answered the item incorrectly, so we asked the participating teacher three crucial questions for informing instructional practice: *What might the student be thinking? What question would you ask to clarify the student's thinking?* and *Would you take any additional instructional steps to remediate the student's work? If so, what?*

While we believe that the Data Analysis Scenario may provide important information on the ways in which classroom teachers analyze and interpret assessment results, we also realize its potential limitations. Foremost among these is the fact that since the assessment results are fictional, teachers are unable to bring contextual knowledge to bear on interpreting results for individual students. For example, with a hypothetical set of results, a teacher cannot attribute low performance to potentially contributory factors such as the student's language status, health status, or other disciplinary or familial problems that occur in real life. For this reason, we believe that the Data Analysis Scenario is best used in conjunction with a semistructured interview with the teacher about her own assessment results. This method is described in more detail in the following section.

Figure 1: Interim Assessment Results Spreadsheet for Philadelphia



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.

Winter interviews. The winter teacher interviews consisted of questions focused on planning for and teaching during the allotted reteaching days. Many of these questions attempted to link teacher behavior observed during our classroom visits with teachers' intentions and with teacher use of assessment information. We also asked about professional development opportunities available to teachers since the first round of interviews and about other potential supports for interim assessment use. During this round we were particularly interested in the technological features of Philadelphia's Information Management System that teachers use. Finally, we asked participating teachers if the term "formative assessment" was one with which they were familiar.

As part of this interview, we asked teachers to bring copies of their most recent interim assessment results with them. We asked both about class-wide patterns of performance as well as about mathematical concepts that seemed to present difficulty for students. These questions were designed to closely mirror the previous questions on the Data Analysis Scenario. In this way, we hoped to get a more complete picture of teachers' individual approach to analyzing interim assessment results. During these interviews, we also noted that some teachers had taken extra steps to organize their data beyond the ways in which the district presents this information, such as by writing the students' names next to items that they answered incorrectly. As we continue to analyze our data, we will make note of ways in which teachers modify the presentation of assessment results to aid interpretation.

Spring interviews. The Spring teacher interviews gave us an opportunity to confirm trends in teacher formative assessment use that we had begun to identify. Specifically, we sought to further explore teacher use of interim assessment results to understand student thinking and to help identify their own professional development needs. We also used this final round of interviews to ask teachers about the role of classroom assessments in light of the annual state assessment (the PSSA) that had just been administered.

In order to gain a broader and deeper understanding of teachers' use of interim assessment results, we linked several questions to two types of artifacts in our spring interviews: (a) an item from the most recent interim assessment and (b) the Benchmark Data Analysis Protocol, a two-page, district-created analysis and reflection worksheet. We chose one item from each of the most recent 3rd and 5th grade interim assessments by selecting those that we felt offered teachers the most opportunity to learn about student understanding of mathematical concepts. The two members of our research team who had previously taught *Everyday Mathematics* (one of whom had taught in Philadelphia) participated in this item vetting. To our surprise, it was actually difficult to identify items for which the distracters offered meaningful information about student learning. Of 20 items on each assessment, only 2-3 were identified as potentially informing knowledge about student understanding relative to a learning goal (as opposed to merely indicating, for example, that a student could or could not perform a procedure). We then chose one item from these 2-3 based on the relative curricular importance of the mathematical content contained therein (e.g., operations were given precedence over measurement) and on the perceived difficulty of the item. Much as we had done in the Fall and Winter, we asked teachers to describe what students who got this item incorrect might have been

thinking, what steps the teachers would take to confirm or disconfirm this hypothesis, and how they might address student misunderstanding.

As mentioned above, we also asked several questions about the Benchmark Data Analysis Protocol. While completing this worksheet is officially voluntary, in the Spring 2006 background interviews, all of our participating principals reported that they expected their teachers to complete these forms and hand them in to the principal. Our particular interest in the Benchmark Data Analysis Protocol was whether or not teachers used it to report their own professional development needs, and whether or not it assisted them in analyzing student results.

Observation of District and School Meetings

In addition to the information that classroom observations provided about how teachers used interim assessment results, we were interested in how schools and districts organized and used the information from these assessments. In order to gain a more complete understanding of assessment creation and data use at the school and district levels, we attended principal meetings in Philadelphia (so-called SchoolStat meetings at which several types of "performance indicators" were discussed). We also observed several professional development workshops for both teachers and principals. Finally, because grade-level collaboration was mentioned as a possible support for teacher interim assessment use, we also attended selected grade group meetings in our sample schools. In each case, we took field notes, and used this information to triangulate our findings from the teacher and principal interviews and to help contextualize our findings from teacher observations and interviews.

Collection of Relevant Artifacts

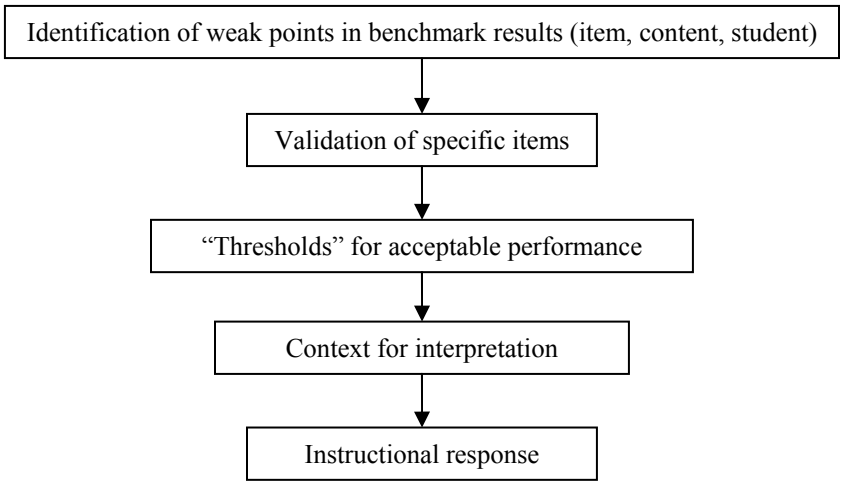
Since this is a study of assessment use, we collected copies of all 3rd and 5th grade interim assessments in mathematics given in the 2006-07 school year. We also purchased the 3rd and 5th grade *Everyday Mathematics* program in order to better understand the learning goals that were to be achieved during this study. We acquired copies of both the Philadelphia School District's pacing guides and of any additional districtwide assessments that were made available to us. In many cases, we were also able to collect examples of teacher classroom assessments. In some cases, teachers offered blinded examples of student work on the interim assessments. As mentioned above, a few teachers had constructed their own data organization templates, and, when possible, we collected copies of these as well.

IV. Teacher processes of data analysis and interpretation

Analyzing benchmark assessments was a universal practice among the teachers in our sample. At a minimum, all of the teachers had experience looking at printouts of students' benchmark results, and most were comfortable accessing those results on SchoolNet (see below). Additionally, nearly all of the teachers interviewed demonstrated the ability to link analysis of individual items to curricular content areas and/or Philadelphia standards. In the case of the latter, use of SchoolNet appeared to facilitate this process.

Figure 2 illustrates the steps most commonly taken by teachers when looking at their benchmark results. In nearly all cases, teachers begin by identifying weak points in their class' performance, either focusing items or content that proved challenging and then moving on to individual students or vice versa. To better understand these weak points, teachers often validated specific items to ensure that they accurately reflected their students' mathematical understanding. Whether or not a particular result was directly responded to depended on the personal "thresholds" for acceptable performance that were embedded—sometimes unconsciously—in teachers' analyses. These thresholds varied considerably and were influenced by a variety of contextual factors such as past student performance, teacher background knowledge, or position of specific content within the district's curriculum or pacing cycle. These same factors played an important role in shaping teachers' overall impression of benchmark results, which in turn directly informed whether and how they responded instructionally.

Figure 2: Teacher process for analyzing, interpreting, and acting on benchmark results



Analysis processes

In the Fall of 2006, 29 Philadelphia teachers were presented with the Data Scenario, and were asked to walk the researcher through the process they would use to interpret the data. Transcripts of the interviews were coded to capture the sequence of analysis steps employed by teachers.

There was considerable uniformity in the initial steps teachers took in analyzing benchmark data. First, nearly all of the teachers (86%) started by looking for weaknesses. (The rest began by reviewing and assessing the overall performance of the class.) Of all teachers, the majority (59%) began by looking for weak content areas, either by looking directly at the standard to which an item corresponded or by identifying individual items that covered the same curricular content. Roughly 28% of teachers began by looking for individual low-performing students rather than at content issues.

Regardless of sequence, nearly all teachers in the sample (86%) eventually looked at individual items, content, and students when analyzing the benchmark data. Worth noting was the common pattern, demonstrated by 79% of teachers, of moving from analysis of items to content areas/standards or vice versa. In other words, teachers had little difficulty linking items, content, and standards. This suggests that at least on a superficial level, the vast majority of teachers were analyzing benchmark data in a manner consistent with school district expectations: identifying weaknesses in student performance and relating those weaknesses back to instructional content.

Use of Information Management Systems

All of the teachers in our sample were familiar with SchoolNet and had used it at some point, though some were more familiar with specific features of it than others. Generally, teachers used SchoolNet to facilitate their analysis in two ways. First, they were able to toggle between the item analysis for their class and the specific problems from the benchmark test. This allowed them to assess the validity of the question itself (see below), and to try and make sense of students' incorrect responses.

> Let's see, for number one, a lot of them put B instead of…A. And this is why I would…go and look at, there's a place that you can click on for the question to come up to view the problem. And so, I could look to see what exactly the problem was, and…because it's geared to Everyday Math, I might be able to see that little thing that they didn't get, the way it was set up, it could have possibly been that more than just the computation of the problem itself. So, I would probably go and look at that.

Second, teachers used SchoolNet to link benchmark items to Philadelphia content standards. Several respondents reported that viewing the data by standard (rather than by item), allowed them to more easily hone in on the specific content they needed to cover during the reteaching weeks.

> This is on Align. And then also on the align tab there is each of the standards, and it gives you a standard mastery by individual benchmark, so then I can take just a real quick look and say, okay, well I know that the most people need work on, I guess right here, whatever this is, skip counting, which you think that they would know. So that would definitely be something that we needed to work on, was the skip counting. So this is another way.

> You see, these tell me right here. These are the [items] that the kids really had trouble with and the content and then all of the different standards. So that helps me so that I know what to reteach. I am not going to re-teach something that everybody got the question—you know, if they got it right, I am not going to reteach that.

Use of SchoolNet for more complex tasks, such as generating supplemental assessments or identifying curriculum, was far less common among teachers in our sample. Two teachers specifically referred to using SchoolNet to link back to relevant Everyday Math units, and just one discussed generating worksheets based on student benchmark performance. Still, teachers were generally highly complimentary of SchoolNet, noting that it both saved them time and helped them focus on students and content areas that most needed their attention.

> I think it's really good. I really appreciate having data and having the numbers just done for me. A lot of times, as I alluded to earlier, I have an idea of where they're struggling, but to actually see it numerically and get it instantly, it saves me time and it really helps a lot. I think it's terrific.
>
> .

Validation of test items

In analyzing the benchmark assessment results, we also found that some teachers went back to the actual assessment items to determine the quality of information that could be gleaned from student performance on individual items. Some type of validity check was conducted by at least half of our teachers, although these actions were not referred to using traditional psychometric jargon. Rather, several teachers mentioned that after looking at students' performance on the whole test or by standard, they would go through the most problematic items one by one.

Of the teachers who mentioned conducting validity checks, the most common question was the degree to which an item (or set of items) assessed students' true understanding of mathematical concepts and procedures. Teachers spoke of two main threats to validity. First, teachers mentioned that they were far less able to make use of the benchmark assessment results when the content tested had not yet been taught. During the year of our study, this occurred for two reasons: because instruction had fallen behind the Planning and Scheduling Timeline and because teachers perceived that items had featured content that was not to have been taught yet. For example, several teachers expressed concern about the first item of the 5th grade January benchmark assessment, which required students to multiply two fractions.

> We looked at this and flipped. What I did was I just said—I did it quickly—"Whenever you have 4 x 4, I mean four 4's, when the numerator…" I had already taught that, numerator and denominator are the same equals one whole. "What is one whole? One times anything, so 1 x 3/11?" That is not the way you teach multiplication, but I was able to do it for this. So that was a problem. We looked at that and could not believe that this was the very first problem and it was not even in our curriculum at this point. (5th grade teacher)

The second main validity concern that teachers expressed was about item construction. Several teachers mentioned that the language used in particular items confounded their ability to use the benchmark assessments to judge their students' math knowledge, prompting one 3rd grade teacher to ask, "Are we measuring reading or are we measuring math?" These teachers

mentioned vocabulary as the main issue—either new mathematical vocabulary (e.g., probability versus percentage) or vocabulary that may not be part of the everyday experience of their students. In discussing how her class performed on the January benchmark assessment, a 3rd grade teacher considered:

> One thing that became clear to me is that the language in some of the word problems was difficult for them, and that might have been a hang-up. Like with the greeting cards, if you were observant, you saw that no one in that group knew what a greeting card was except for one girl. It was a math problem about a box of greeting cards.

In sum, item validation was used as a first check on class-level benchmark results that were unexpectedly low, or appeared otherwise anomalous. It is worth noting, however, that questions about validity were almost never raised when students performed *well* on an item. As discussed later, in those instances teachers almost always assumed that their students had mastered the content and focused on the areas in which they struggled.

Teacher interpretations of benchmark data

Teachers' processes for interpreting benchmark data were profoundly influenced by a variety of factors, including their knowledge about specific students' background or past performance, student performance in relation to their peers, the position of the specific benchmark (or item) within the school district's curriculum and pacing cycle, or teacher perceptions about which mathematical content was especially challenging for students. These factors contributed to the development of teacher "thresholds"—criteria for determining whether student performance required an instructional response during the reteaching week—and also shaped their overall evaluation of their students' progress.

Individual "thresholds" for benchmark assessments

Teachers' interpretations of benchmark results revealed the existence of personal "thresholds" that profoundly influenced their interpretation of the data. That is, teachers looked for a minimum score on the mathematics benchmark that, to them, indicated whether their students had mastered the concepts introduced during the previous five weeks. These personal thresholds were individually defined by teachers and influenced by their knowledge of their students and their abilities, as well as by teachers' beliefs about content difficulty and how children learn mathematics. These thresholds appeared to vary by student, by class, by time (when during the year the benchmark was given), and by range of students' responses on the benchmarks.

In one school, teachers referenced a "green," "yellow" and "red" system whereby a green indicated a score of 85-100 ("mastery" or "proficient"), yellow indicated scores between 65 and 80 ("strategic"), and red scores of 65 or less were considered "at risk." Yet, even at this school, teachers appeared to construct their own personal thresholds. Across all the schools, teachers frequently used the terms "advanced" and "proficient" when discussing their students' scores on the mathematics benchmarks. And while teachers' thresholds might also have been influenced by

the these more fixed distinctions, their own thresholds appeared more mutable and fluid; they could fluctuate up or down depending on content, context, student, and even from benchmark to benchmark.

It is worth noting that the term "threshold" was introduced by the research team; it was not native to the teacher lexicon. We used the term in questions included in the data scenario:

- *Are there any students who appear to have mastered the material? …What would you consider the "threshold" for mastery…?*

- *Probe: Are there any students who appear to having trouble with this material? …What would you consider the "threshold" for recognizing a child as having difficulty?*

Teachers had ready responses to these questions. It was clear that teacher thresholds were not randomly selected but were shaped by several factors. In this section we examine teachers' individual thresholds, their significance, and the factors that informed them.

The significance of teachers' thresholds. On paper, five dedicated days to reteach and remediate to some students, and extend concepts to other students, is luxurious. And relative to other school districts, such an instruction and assessment schedule is remarkable. On the ground in urban classrooms, the reality loses some of its luster. Teachers must pick and choose how they will spend these five days and with whom in a class of 30-33 students whose needs vary considerably. Thresholds help teachers make hard decisions about where and on whom to focus, what to emphasize, and what and whom to ignore. Ideally, a student's score on the latest mathematics benchmark would determine and shape his/her instructional experiences in the last week of the district's teaching and assessment cycle. In reality, however, many challenges can conspire to undermine the meaningfulness of the "reteaching" week for individual students.

Given the pressure on Philadelphia teachers to raise their students' achievement and test scores, the bulk of some teachers' attention during the 6th week was often directed at students hovering near proficiency. With limited resources, there was some evidence that teachers tended not to direct their instructional attention to those students at the far reaches of the scoring spectrum—those students scoring at the extreme low or high end. A 3rd grade teacher's comments reflected this practice of targeting the near-proficiency students:

> I would look at the kids who are in the 70s and think, "Who among this group are ones I'm missing an opportunity to yank up?"

A 5th grade teacher revealed how she had altered her personal threshold to capture more students in the proficiency category:

> He ends up getting a 75 where I know he could have probably got a 90, but you know what? …I have changed my whole attitude. Before, 75 was never acceptable to me. Now, I am thinking, you know what? It is one more kid I am going to have in the proficient category. (5th grade teacher)

Used in this way, personal thresholds were enormously important in guiding teachers' activities and directing their focus during the reteaching week.

Threshold ranges. The 3rd and 5th grade Mathematics Benchmark Assessment is a multiple-choice measurement containing 20 questions. Teachers' personal thresholds for mastery varied considerably, but for most teachers the marker fell between 60% and 80%.

> I always look at 80% or above, which really means that they understand it. …if they're really having trouble, I would say below 60, because they're not passing at all. The ones who are in between are, like, average. They're kind of getting it, but maybe still having problems. (3rd grade teacher)

> I would like 75 or higher… [70] is still borderline to me, so that's not enough, not giving me enough. (3rd grade teacher)

> I think if they're 70, that's not good enough. … Seventy, to me, means you're just getting by, by the skin of your teeth. (3rd grade teacher)

> I would say any one with…80% or higher [has mastery]. And the kids at …70 and 75, are making progress. (5th grade teacher)

Student-specific thresholds. Many teachers maintained a sliding threshold of sorts, adjusting their personal thresholds depending on the student. For example, when a 3rd grade teacher was asked to explain her use of the term "good job" on the mathematics benchmark, she said that while she personally "like[s] the 80% or above, depending on the child, if they got a 70%," she would be satisfied. In interpreting their students' scores, many teachers relied on their background knowledge of individual students.

> If that [student's benchmark score] seems in line with what I know that the student can do, then I'm happy. And if it's not, if I have a student here who's done, like, 70 or something, then that's kind of where my focus would be. I'd hone in right there and figure out, "Well, what did he or she do wrong? Normally an A student, [gets a] 70%? What's going on?" And then try to figure out what [happened]. (5th grade teacher)

> I would look at individual students. And, of course, I would have already made my classroom observations and all through my unit. And I pretty much know where I expect my higher students to be scoring, where I expect my middle students, and so on. … But if I saw my higher students, and they're scoring down in the 60th percentile, then I would know—send up a little red flag that's something's not right here. (5th grade teacher)

Because teachers often possessed deep and detailed background knowledge of their students, many expressed surprise when a student's score on the benchmark was inconsistent with what they knew about a particular student:

> They were really getting into this [*Everyday Mathematics*] Unit. And they seemed to have gotten what was being asked of them. So, I expected to see most of them at least in the 70, 75 or higher. One of the students was really low…he was like a 55%. That surprised me because I know he's having problems with math, but he also has an outside tutor. And I thought he was beginning to get it. (5th grade teacher).

Class-level thresholds. Teachers' personal thresholds were also informed by the scores of the class as a whole, similar to grading on a curve. When a large majority of students had scored well on a particular mathematics benchmark, teachers' thresholds were upwardly adjusted. In these classrooms, teachers' attention during the reteaching week targeted the proficient students in an effort to bump them from "proficient" to "advanced."

> Well, the average [mathematics benchmark] score in this class is 82%. Certainly, the kids who have 100s are very secure. Ninety-five percent, that was just one wrong. Also, 90 is pretty strong. But I wouldn't necessarily say "mastery" because my goal is really to pick up each kid as high as they can. So, 95 and above is considered advanced, and from between 80-94 is considered proficient. So, if a child is proficient, my goal is to try to help them reach advanced. So, I wouldn't rest on my laurels or allow them to rest on theirs with a 90. (3rd grade teacher)

Yet another teacher had a different expectation for her class, and when the class did not meet her "target," she adjusted her threshold downward:

> I was hoping for 70% average and they had 64. So, I was happy with that. They're progressing, which obviously is a good thing. So, I was happy with that. But I guess I was hoping that they did a little better. But they are doing well, so I am happy with that (5th grade teacher).

Here, the teacher's threshold adjustment is explained by her students "progressing." Thus, in settling for a lower class average, this teacher's higher expectations for her class are tempered by her satisfaction that the students are making progress.

Factors influencing personal thresholds. Teachers' personal thresholds were also influenced by the school district curriculum and pacing schedule. Some teachers expressed different expectations for the first benchmark of the year, given in October, than they did for the March benchmark. As such, these teachers were less concerned with lower scores on the October benchmark than on a benchmark given later in the school year. A 5th grade teacher commented that she's "fairly satisfied with a 65" on the first mathematics benchmark administered in the school year as "it's…been a summer away from it." However, this same teacher maintained that there are certain basic mathematical skills that students should possess *regardless* of when the benchmark is administered:

> And at the same time, there are always certain things that I feel on the first unit that they really should do well in, because the first marking period, obviously, is generally…a review. So they really should. Some of these skills that they're seeing on the [benchmark] are basic addition problems. And if I see somebody that gets that wrong, I have to question whether or not it was just a silly mistake. But I would look at it and say, "They got this wrong. Let me see if I can just take this person, one on one, and make sure that there's nothing really going on."

Teachers' knowledge of the district's mathematics curriculum, *Everyday Mathematics (EDM)*, also appeared to help shape and determine their personal thresholds. *Everyday Mathematics* is a spirally structured program and students receive ongoing opportunities to review and practice skills and concepts after they are first introduced. Because different skills and concepts are introduced at different times, the second edition of *EDM* recognized and distinguished between "beginning," "developing," and "secure" skills. In discussing their personal thresholds, some teachers expressed less concern for a lower score on a "beginning" and "developing" skill than on a "secure" skill:

> I would say maybe about 75% of them…got it. …And plus, I don't think this was a secure goal at this time. So, since this is a spiraling program, all of them weren't supposed to be able to master it. (5th grade teacher)

These so-called "beginning" skills were recently introduced concepts that the teacher had not devoted a lot of instructional time on during the five weeks that preceded a particular mathematics benchmark. These *EDM* distinctions—"beginning," "developing," and "secure"—coupled with the spiraling nature of *EDM* indicated to the teacher (and the students) that the mastery was not expected at this time. However, the mathematics benchmarks make no such distinctions. It was the teachers' knowledge of *EDM*, of the curriculum's scope and sequence, and of the pacing guide, that helped them determine how much "weight" and what threshold to set on particular benchmark items.

Context for interpretation

The same set of factors that influenced teachers' personal thresholds also colored their overall interpretation of student performance. Teachers interpreted student benchmark scores in the context of their expectations, both for individual students and for the class as a whole. As discussed in a previous analysis (Oláh et al., 2007), teachers frequently used the benchmarks to validate their impressions of student strengths and weaknesses based on other assessments, performance on previous benchmarks, informal observations, or nonacademic background information. According to one 3rd grade teacher:

> I can't say [benchmark results are] a big surprise, because as we're going through 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're got them, if they're understanding it.

In the following exchange, a 5th grade teacher reflects on being "surprised" by the results of the benchmark assessments:

> **Q**: If you look at these results from the last thing, is there anything on here that kind of surprises you or jumps out at you?
> **A**: Yes. I actually had one student that did surprisingly well compared to how she has done in the past. So that was something going through this data that really jumped out at me.

Below, a 3rd grade teacher explains a child's poor performance on the benchmark assessment:

> This child is the only child that did poorly in my class, but basically because he doesn't come to school until 10:00 and we teach math in the morning. So he misses math everyday. So obviously, it's not a learning problem. It's a not showing up to school problem.

Teachers' interpretations of benchmark results were also influenced by their expectations for the performance of the class as a whole. As in the case below, these expectations were informed by teacher understanding of content areas with which students struggle.

> I noticed some of my children do not have a strong place value background. And they should be able to write numbers…and read numbers to the billions. And they're having difficulty with that. So, I am alerted already that I definitely need to do some re-teaching with that, where some of the prerequisite skills that I assumed they had with reading numbers are not there.

Teacher expectations for whole-class performance were also rooted in the organization of the curriculum. As noted above, *Everyday Mathematics* is organized around "beginning, developing, and secure" skills, with secure denoting the highest level of student mastery. Teachers used these criteria to determine the level of acceptability of student performance; low performance in a secure area alarmed many teachers.

> One of the questions that they got wrong surprised me too…The question about the expanded form with decimals…That was something that I really didn't expect them to have a hard time with, but…six of them got that wrong and I thought that was one that was pretty obvious…They should be secure, right, and that surprised me that they were not.

Finally, curriculum pacing factored into teacher expectations for class performance. Skill areas that teachers had ample time to cover were expected to be more secure than those they had touched on briefly, or not at all.

> I was surprised to see the multiplication was so high. I was glad, of course. But 19 out of 25 kids, that's pretty good, because we just started covering multiplication and division.

> This is only my second year in 5th [grade]. So, I know 4th grade curriculum. And I know what I taught my children, so I expect the incoming children to have those prerequisite skills, especially in geometry and things. So, if I show you an acute angle, you're supposed to know what it is from 4th grade. And if you don't, there's a red flag in my head. "Uh-oh. I'm going to have to do a lot of introducing of certain things they should have been exposed to already." (5th grade teacher)

Diagnosis of student understanding and misconceptions

A crucial question about teacher analysis of benchmark assessment scores concerns any deeper analysis that teachers do once they have looked at overall patterns of scores. In order to investigate the types of "diagnoses" that teachers perform, we interviewed all of our teachers about both their own assessment results as well as about a select number of items (the misconception probes detailed in the Data Sources section above). In the latter case, the important question we asked of teachers was *What might the student be thinking?* (when the student answered the question incorrectly). We see this moment of analysis as a critical juncture between the reporting of benchmark assessment data and modification of instruction. In subsequent analyses we will seek to explore the relationship between the types of diagnoses made and the types of instructional modification. For the present paper, however, we begin by describing how teachers attributed diagnostic information to individual student performance on specific items in different ways. We recognize that the four categories detailed below may simplify what is, for many teachers, a very complex decision making process, and we do not claim that these categories are mutually exclusive. In fact, teachers may attribute student performance to multiple factors simultaneously or the difference between some categories may not be as discrete as researchers have assumed it to be (cf. Baroody, Feil, & Johnson, 2007). We therefore view this analysis as a starting point for further inquiry.

By far the most common diagnosis of student error on actual benchmark assessment items fell into what we call the *procedural* category. Diagnoses of this type focused on missteps in applying algorithms or on computational error. Over half of teacher diagnoses included some kind of procedural diagnosis; students were seen to have particular difficulty with items that required multiple steps to reach an answer. For example, one 3rd grade teacher, considering her students' performance on the January benchmark assessment, commented that "doing the double-digit subtraction problems with regrouping, that was the most problematic, I thought, because they were still having trouble with that process of doing the regrouping."

A less frequently mentioned set of diagnoses fell into a *conceptual* category, in which teachers mentioned problems in students understanding basic definitions or more complex ideas. For instance, when speaking about their own class' results, 3rd grade teachers mentioned that items featuring place value were some of the most difficult for students, while 5th grade teachers pointed to fractions as the one subcontent area that the benchmark assessments drew attention to. One 5th grade teacher explained her interpretation of some students' responses to a fraction identification item:

> I remember there was one question I had four boxes and the first three were shaded in, and the last one, it didn't have individual boxes inside shaded in. It just

had three-fourths of it. And I think some of the students thought—I don't think they put together that each one of those [the big boxes] could be divided up into four, so the denominator would have been 20 because there were four in each of the five boxes. They were thinking of them as wholes.

A few teachers mentioned that word problems also had the potential to pose conceptual problems for students in that students must know about different algorithms and be able to choose the correct one to apply.

Many teachers also attributed student errors to *other cognitive* weaknesses. These included a list of possible causes for student underperformance, including, but not limited to, weak reading ability, difficulty maintaining attention, and English language proficiency. As might be expected, errors on word problems and on multistep procedural problems most frequently elicited this type of diagnosis. For example, a 3rd grade teacher in a school with a high proportion of ELL students believed that a subtraction word problem that ended with the words "how many more marbles does he need?" had posed difficulty because when her students saw the word "more" they summed the minuend and subtrahend instead of subtracting the latter from the former. She believed that her students "just say, Oh, 'more', altogether, let's add." Although our questions focused on teacher response to student error, one 5th grade teacher attributed a student's superior performance to increased attention to task in that, as the teacher explained, "he usually doesn't do quite that well….it goes to show you what he can do when he is paying attention, because he did exceptionally well."

Finally, teachers also offered *contextual or external* diagnoses, according to which student mathematical performance fell short due to factors that were seen to be outside of the teacher's or school's realm of influence. These tended to consist of perceived distal causes of the other proximal diagnoses. For example, several teachers mentioned students' lack of background knowledge as contributing to difficulties in comprehending word problems. A teacher who taught two classes of mathematics mentioned that one class was "calmer" than the other class, giving all students the opportunity to "get more into the work … [taking] more time to look things over." We are very interested in examining teacher planning in response to contextual/external diagnoses since it seems that a teacher's concept of his or her role in facing this type of challenge can vary greatly. For example, some teachers may use diagnoses of this type to demonstrate the lack of influence that their instruction can have on student performance, while other teachers may believe that it is primarily because of these external obstacles that they must try even harder to increase their students' learning. We will also continue to examine teachers' differential responses to student performance on actual benchmark assessment items in comparison with the "misconception probes" that we administered to see if items designed to assess student conceptual knowledge do in fact elicit more conceptual diagnoses than items that emphasize procedural skill.

VI. Instructional response to benchmark data

Philadelphia teachers appear to have some latitude in planning their lessons and activities during the 6th week of the district's instruction and assessment cycle, the "reteaching week." The district's expectations for how teachers should address their instruction are guided, at least on

paper, by the Benchmark Data Analysis Protocol. The district-created protocol directs teachers' attention to the item analysis feature of SchoolNet. The protocol prompts teachers to navigate SchoolNet and respond to the following:

- Using the item analysis report, identify the weakest skills/concept 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?

The second part of the protocol, entitled "Teacher's Reflection," is intended to be completed by teachers after every benchmark. One of the five reflection prompts expects teachers to address how they are going to instructionally respond to students:

- In order to effectively differentiate (remediate and enrich), I need to…

Although the district did not require teachers to complete and turn in the protocols to their principals, the teachers in our study reported that it was required. At the same time, teachers noted that there was little monitoring and/or followup to these protocols.

Beyond the Benchmark Data Analysis Protocol, there seemed to be little guidance for teachers about how to act upon their analyses of benchmark data. Still, it appeared that many 3rd and 5th grade teachers adopted common instructional responses and strategies. Below, we detail teachers' instructional responses to the benchmarks and their approaches taken during the so-called reteaching week.

A "triage" approach

During the reteaching week, 3rd and 5th grade teachers generally seemed to follow a "triage" approach to instructional planning on the basis of benchmark assessment results, devoting the greatest amount of time and effort to those students and content areas that most urgently required their attention. A 5th grade teacher succinctly summed up this approach saying, "I can't reteach every single thing." Using their personal thresholds as barometers for their students' mathematical mastery and understanding, teachers decided whom to target and what to emphasize during the five days that followed the administration and scoring of the benchmark. In analyzing the benchmarks, many teachers looked for particular items that gave the class trouble and also determined if it was challenging for just a few students or for many. According to a 3rd grade teacher:

> If it's half the class…I'll just reteach the whole thing. But if it's a few children…, then I would definitively pull them out and get some special homework for them to work on.

In general, teachers targeted the lower performing students and also those content areas that proved the most problematic for students. Or, put another way, in the words of a 5th grade teacher, "I'm not going to waste a whole lesson reteaching something that 90% of the students got. That's just not beneficial for the other students." Many teachers described a similar approach:

> A lot of kids got the same ones wrong. Like, for example, [item] five. There's a lot of kids who got 5 wrong. And a lot of kids who got [item] 14 wrong. So, then I go back and I see, "Well, what was that question and what was it that the question was asking?" … So, then, I would take a look at that and see, "OK, well, I need to reteach that."

> Well, the things [items] where the majority of the class got it wrong, like those 60% that we discussed earlier. I would immediately go back and revisit those. And I might even retest them myself with the teacher-made assessment to see…if there was any improvement.

At the same time, many teachers took note of what they apparently had taught well and, based on the results of the benchmark, that their students had understood. According to one 5th grade teacher, "OK, [items] six and seven look good…So, these two tell me that they're pretty solid on this. So, this isn't something that I necessarily have to go over."

As noted earlier, decisions about whom or what to "triage" depended on teachers' interpretation mechanisms. While low-scored content areas were almost always addressed, decisions about reteaching both students and content were shaped by teacher expectations, background knowledge of students, and performance relative to curriculum and pacing. The ways that teachers analyzed data thus informed directly their decisions about how to use instructional time during the reteaching week.

High-scoring students

Overall, there seemed to be less *direct* instructional attention given to students who had scored high on the mathematics benchmarks. While teachers mentioned their high scorers in interviews, in planning for the reteaching week their focus was on the students who had not done well. "Enrichment" for high-scoring students often consisted of short-term activities, extra-worksheets, more Everyday Math game times, and time on the computer. There was evidence from both 3rd and 5th grades that high-scoring students received less direct instruction during the district's reteaching week. A 5th grade teacher remarked:

> I don't want to say [high-scoring students] get busy work, but they would be the ones who I might give an independent or a small group project to do, creating a graph. *Everyday Math*, our math series, has games. They're…educational games.

Similarly, a 3rd grade teacher commented:

> The ones that did a little bit better, maybe…bump them up and give them some enrichment while I work with the ones that need help.

As we noted earlier, teachers often turned to their high scorers for instructional support in the form of peer teaching. It was not uncommon for these higher scoring students to be paired with their lower performing peers during the reteaching week. As one 5th grader teacher noted, "[Students scoring in] the 50s and 65, I think I would definitely have them working maybe with the higher students as peer tutoring." A 3rd grade teacher shared a similar strategy when asked about her high-scoring students during the reteaching week:

> I would make sure that the kids, these kids who were low, were seated next to children who were strong and whom I feel could help them and they could have a good working relationship with. So, I'd rearrange the seats accordingly.

Instructional responses

Philadelphia teachers used a package of whole-group instruction, plus small group, and peer teaching, as grouping techniques when redressing content during the reteaching week. As they are given a full week, teachers might employ these different strategies at different times during the week or even within a single mathematics lesson. Not surprisingly, they tended to respond to more widespread errors with whole-class instruction. A 3rd grade teacher spoke of this particular practice, saying, "If almost half the class doesn't have that, that's something that needs to be reviewed as a whole-class thing." Presented with the Data Scenario, a 3rd grade teacher observed:

> I notice that question 8 looks like…one they struggled with, and question 4. It seems like about half [of the students]. …So, I'll take this particular question that maybe half the class struggled with [and re-teach to the entire class].

Conversely, teachers appeared to favor small-group instruction when the benchmark errors were less common among students. Given the challenges often posed by meeting with small groups during the regularly scheduled math lesson, many teachers found time to meet with students outside of math class. One 5th grade teacher described how she might meet with students needing additional instruction during the reteaching week, and how she would link this small group instruction with the students' experiences during class:

> Probably, what I would do would be to ask them to come in the morning a little bit early, because they're here early enough. And I'm here all the time early. So, for them to come—and maybe come with somebody else, have a couple at a time come—and to work with them like that. You can get a lot done in a very short time with that intensive kind of thing. And then I would just kind of, like, keep an eye on them and if I—when we're talking about the topic, I would kind of help them build their confidence in their ability to answer these questions by calling on them when I knew that they knew the answers to these similar topics. But basically work with them one on one and just kind of like touch base again. I just keep checking in and seeing how—"Are you OK with this? Do you understand

> it?" …So, then I had an opportunity to have them come in in the morning and sit and work. And then they were able to get it when there was no pressure.

A 3rd grade teacher shared her small-group practices:

> I pull small groups in during recess. And I try not to take their recess. I might just take a couple of minutes. I want everybody to go to the board—"I want you to do two problems and then you can go out for recess."

A handful of teachers had student teachers assigned to their classrooms during the course of our study. Having a second adult in the classroom allowed teachers to keep providing instruction to most of the class while the student teacher sat with a small group of children in the back of the room, providing them with extra instructional support. A 5th grade teacher described how she used a student teacher to provide one-on-one instruction to a student having difficulty:

> And now I have a student teacher, so I can have her work with students that really just are not getting this. What we are working on today, one student just was not getting it at all. So I sent her into the hallway with the student teacher and they just worked on it. And then I'm in here, working with the others.

Another 5th grade teacher reported using her "two brightest kids" in those "in-between times" when she did not have a student teacher: "They do help me tutor, especially when I do not have a student teacher."

In two of the schools, we observed regular volunteers in the classroom. These volunteers, often retired citizens who helped out in the same scheduled classrooms for as much as four hours a day, 3-4 times a week, would also work with a small group of children needing additional instructional help.

An additional strategy, often used to supplement whole group and small group instruction, was peer teaching. In part, we suspect that peer teaching proved such a popular instructional strategy for teachers because of limited resources. That is, many teachers were pretty much "on their own" in their classrooms and relied on stronger students to help teach the students who had performed less well on the benchmarks.

When a 3rd grade teacher was asked how she would reteach a particular concept, she replied:

> This is where I would partner up the students that understood it versus the students that didn't understand it, and they can share their strategy, because I always tell the kids that we all learn from each other, and that someone else's strategy might help you get the answer better. So, this is where I would use the different groups and partners, where I might match them up and someone who got a 95 versus somebody who got…a 60 can maybe work with them to think of what they were thinking and go over it.

A 5th grade teacher described a similar approach:

> A lot of times I'll get another student to help with that, because a lot of times students are better with other students. And if you get a student who's really good at letting another student learn, not to show them, "This is the answer. Write that down," but explain…I have a couple students that are really good at that, with explaining stuff.

At least one teacher expressed concerns about peer teaching, cautioning that "buddy work is fine, but sometimes it doesn't work. The other kids don't want to do it, they're tired of doing it, they're tired of helping."

Individual remediation

There was very little one-on-one instructional remediation reported during the reteaching week. A more common instructional response was to cluster low-performing students into small groups for a limited amount of time, either during math class or, more commonly, outside of class. Yet, while not as common, some teachers noted that they tried to meet briefly one-on-one with students in a "student conference" during the class period. As described earlier, teachers also reported working with children outside the regular class period—at lunch, during recess, or before school. A lack of resources and personnel made it challenging for a teachers deliver one-on-one instruction to individual students. In response to these persistent shortages, teachers had to be creative in how they used their higher scoring students, their student teachers, and classroom volunteers.

Procedural emphasis

Just as teachers' diagnoses emphasized procedural challenges, teachers' reteaching activities appeared to focus first on retracing procedural and correcting procedural steps. A 3rd grade teacher described how she would focus on "step by step" procedures and also on test-taking strategies:

> Tell them to look for, like, key words and clue words and things like that. Underline and pull out your information. And a lot of time they just…they add it up. They're not reading what the question is asking. So, that's another big thing that I take my time and teach…step by step. …It could take us a half hour to do one problem because I make sure that they pull out the information.

Another 3rd grade teacher referenced and credited the district's math program, *EDM*, for directing their instructional attention to procedural missteps:

> Another nice thing I like about *Everyday Math* is that they structure so many things and they give you so many nice sheets that you can give the students where they are encouraged to answer a question in a certain way so that, for instance with this kind of a problem, they have a sheet that's set up with a tenths column, a hundredths, a one, and so forth. And I need to see them answer it in a certain way, if for no other reason that, if it simply comes down to a student adding 3 and 4

> incorrectly, I can see that otherwise, they knew exactly which steps to do. And you know, 3 and 4 incorrectly, that's one issue. That could have just been moving quickly. But there's the process in place. And that's why looking at the particular missed answers is so important.

Again, it appears that evident patterns in teachers' analysis of benchmark data (in this case an emphasis on procedural diagnosis) were paralleled in their instructional responses. In tending to students' procedural mistakes first during the reteaching week, teachers themselves appeared to be observing a sequential, step-by-step response.

Changes in instructional practice

Despite this procedural emphasis, analyzing benchmarks appeared to prompt some teachers to adopt new or different instructional practices. Many teachers held the belief that "teaching content another way" would help lower performing students acquire skills and concepts the second time around. A 3rd grade teacher noted:

> I would definitely try a different approach, because, obviously, they didn't get it the first way I did it. Or some kids develop at different stages. So, they might get it the second time I teach it. I would try and do it…a little bit differently.

When another 3rd grade teacher was asked if her teaching would vary during the reteaching week, she responded:

> It depends. If most of the class got it right, got certain questions right, then I would feel that it was a pretty effective way to teach that, and that these children might just need a little extra push, a little more support, to get it. And if they didn't, the reason I would be in a small group with them is to try to find out why that technique didn't work for them, and whether I need to change the vocabulary or the way I'm presenting it, or give them more visual aids and more strategies.

Many times, this "other way" featured the use of visualization or manipulatives, almost as a scripted response. This almost immediate response, across schools and grades, was rather startling to us. Moreover, use of these approaches did not seem to depend on the content being taught, or even the errors that were made but rather, the belief that variety of presentation is beneficial to learning. When a 5th grade teacher was asked how she would correct a misconception about comparing the size of fractions, she responded:

> Different ways of looking at fractions, like maybe cups of water. Maybe not so much 7/12 as maybe going back to just doing ⅓, ½, and ¼, like simply benchmark fractions that they might know. Because ask a kid to give a fraction and they invariably say, "Oh, one-half!" And everything is one-half. That's their idea of a fraction. …Of course, it is a fraction, but they don't really know what that represents. And so what I would do is probably go back to easy ones and start with that and then work up. I would probably try to get them to give me the definition of what that denominator is, and what that really means, and then go

> back and ask them again if they thought that that was—they'd be happy with that part of the pie. I might ask them to draw me a picture of what it is that that they were looking at. "Draw me 7/12 of a pie. Draw all of these and show me what this ate." I might ask another question about how much is one-half of something and three-fourths? I think the pictures would be—kind of let me know. And so if they showed me 12, and then shaded in seven-twelfths, then I'd be really stumped, because then I'd really have to talk to them about it. Because that's a serious—If they could actually represent 7/12 shaded in and all the pies were the same, I really would have to step back and say, "What the heck are they thinking?" and then just maybe go back and do—other than pies—some kind of manipulative. Maybe Hershey bars or arrays or something like that.

Instructional follow-up beyond the re-teaching week

As noted elsewhere in this paper, the spiraling nature of *EDM* guarantees that certain concepts and content will be revisited for additional and more detailed instruction at some point in the future—both within a given school year and from grade to grade. As well, the curriculum spirals from year to year, providing students with more opportunities to learn and work with various concepts. In some ways, the spiraling quality of *EDM* provides teachers with a pass of sorts. That is, teachers well-versed in the curriculum know that students who do not master particular mathematical content in a specified time will have many more opportunities for mastery later on. As such, teachers (and students and parents) realize that if a student does not master a concept the first or even second time it is introduced, there will be other opportunities for mastery.

The district's revamped calendar, allowing for five weeks of direct instruction followed by a week of reteaching, indicates a "hard stop" at the end of the 6th week. In the course of the school year, the six-week cycle repeats itself again and again. A small number of teachers noted that they might give a test or quiz at the end of the 6th week to gauge their students' understanding during the five days in which they retaught. However, there did not appear to be any common or uniform "measurement" that teachers administered to their students. When one 6th week cycle ended (typically on a Friday), a new cycle began the following Monday.

VII. Conclusion

What is profoundly striking in our study of teachers' use of benchmark assessments is just that—*teachers' use*. As we have stated elsewhere, and it bears repeating here, teachers are using the benchmarks. While teachers may not always be using them in the way the district intends them to be used, the fact remains that they are consulting, analyzing, and acting on benchmark results.

Another significant question is *how* the benchmark assessments are being used. Here we find something of a mixed picture. On the one hand, teachers use benchmarks to identify areas of emphasis (both content and students) during the reteaching week, and they are adept at linking

items with state standards and academic content areas. This set of practices and competencies is very much in line with the district's intent. On the other hand, the teachers by and large did not use the benchmarks to make sense of students' conceptual understanding of the content, nor were the benchmarks helpful for diagnosing errors in anything beyond a procedural way. Future analyses will explore the extent to which teachers' conceptual knowledge of mathematics teaching influenced their capacity to use the benchmarks in this manner. What seems clear, however, is that the benchmarks themselves are, as currently constituted, ill-suited to this purpose. In most cases, teachers could learn little from students' incorrect responses to items.

Finally, the tendency to analyze the data in procedural ways was paralleled by a tendency toward procedural instructional responses. The benchmarks appeared to alert teachers to the fact that they needed to "teach differently," but the type of change required did not necessarily relate back to anything teachers learned from the benchmarks. Instead, teachers seemed to draw from a set repertoire of instructional strategies; if one did not work, they simply moved to another. Overall, it appears that the manner in which teachers act on their interpretations of interim assessments was aligned in a broad way with their intended use, but that limitations in their analyses of those data ultimately led to a relatively superficial approach to instructional planning and response.

References

Nabors Oláh, L., Lawrence, N., Goertz, M., Weathers, J., Riggan, M., & Anderson, J. (2007). *Testing to the test? Expectations and supports from interim assessment use.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system: A policy brief.* Washington, DC: The Aspen Institute.

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

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.