2010

Formative Assessment Practices with Benchmark Testing: Phase I

Lisa Abrams
Virginia Commonwealth University, lmabrams@vcu.edu

Angela P. Wetzel
Virginia Commonwealth University, apwetzel@vcu.edu

James H. McMillan
Virginia Commonwealth University, jhmcmill@vcu.ede

Formative Assessment Practices

With Benchmark Testing: Phase 1

Prepared by:

Lisa M. Abrams
Associate Professor
Foundations of Education

Angela P. Wetzel
Graduate Research Assistant
MERC

James H. McMillan
Professor, Chair
Foundations of Education

Virginia Commonwealth University
July 2010

The views expressed in MERC publications are those of individual authors and not necessarily those of the consortium or its members.

Table of Contents

Study Team 1

Introduction 2

Background 2

Methods 8

Results 12

Discussion 40

Conclusion 49

References 50

Appendix A 55
Teacher Focus Group Protocol

Appendix B 58
Data Analysis Codes and Frequencies

Appendix C 62
Code Dictionary

MERC Study Team Members

Lisa Abrams and Jim McMillan, Principal Investigators, Virginia Commonwealth University

Angie Wetzel, MERC Graduate Research Assistant

Kevin Hughes, Chesterfield County Public Schools

Kerry Robinson, Colonial Heights Public Schools

Helen Whitehurst, Henrico County Public Schools

Carole Urbansok-O'Brien, Hanover County Public Schools

Linda Hyslop, Hopewell Public Schools

Janet Covington, Hopewell Public Schools

William Craig, Powhatan Public Schools

Ann Allen, Richmond Public Schools

Introduction

The recent popularity of benchmark testing has become closely tied to formative assessment. Formative assessment has traditionally focused on a process in which evidence elicited from classroom assessments is used to provide feedback to students and inform instructional correctives (McMillan, 2007; 2010, Popham, 2008). However, accountability demands have led to widespread implementation of benchmark, interim, periodic, or quarterly testing at the school or district level that is often labeled "formative." Indeed, some in the testing industry believe that benchmark testing *is* formative assessment. For example, the definition of benchmark testing by the Regional Education Laboratory Mid-Atlantic is that "a benchmark assessment is a formative assessment" (Brown & Coughlin, 2007, p. 2). However, as pointed out by Goertz, Olah, and Riggan (2009), research on the effectiveness of formative assessment has defined it as a practice that is "embedded within classroom instruction" (p. 1). There is very little research that examines how benchmark testing data are used as formative assessment as defined by the classroom assessment literature (Goertz, et al.; Perie, Marion, & Gong, 2009). The purpose of this study was to explore the extent to which teachers' described using benchmark testing data in formative ways, to identify factors that support teachers' use of test results to enhance instruction and improve student learning, as well as to identify barriers that may limit usage.

Background

Tests covering a large amount of material, such as those covering six or more months of learning, would typically be thought of as summative assessments, with some diagnostic information and an ability to identify relative strengths and weaknesses. It is now clear that benchmark testing is administered with the hope that results, which can be quickly returned to teachers and administrators will be used to enhance student learning so that more students will be successful on accountability tests. The goal is to maximize the percentage of students who pass these yearly tests. Such use, however, depends on technical as well as procedural qualities related to the format of benchmark tests, when they are administered, and how results are reported, and most importantly, how the teacher interprets and uses the results to provide individual feedback to students and modifies targeted subsequent instruction focused on student errors, misunderstanding, or lack of knowledge and/or skills.

At issue is whether it is possible to use benchmark testing for what has been carefully and clearly defined as a process or series of steps used in formative assessment (Wiliam & Leahy, 2007; Brookhart, 2007; Popham, 2008). Consider the 2006 definition used by the Council of Chief State School Officers:

> Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes.

Note this definition includes *during instruction, providing feedback*, and *ongoing teaching*. These are characteristics not often associated with large-scale testing. Wiliam and Leahy point out "a 'formative assessment' that predicts which pupils are likely to fail the forthcoming state-mandated test is not formative unless the information from the test can be used to improve the quality of the learning within the system" (2007, p.31). Popham has recently made the same point in his definition of formative assessment, which emphasizes that formative assessment is a "planned process" in which evidence is used so that teachers "adjust their ongoing instructional procedures" or students "adjust their current learning tactics" (p.6). It is assessment with these characteristics that, according to the research, improves student learning (Popham, 2008).

The prevalence of benchmark testing is widespread. In a synthesis of four separate studies related to data-driven decision making, Marsh, Pane and Hamilton (2006) reported 89 percent of school districts in Georgia required some or all schools to administer interim or "progress" tests in mathematics; 50 percent required similar tests in science. One-half of California districts and one-third of districts in Pennsylvania required interim assessments in mathematics. Given the extensive use of benchmark assessments, little is known about how these tests are influencing the instructional process and in turn, student achievement (Marsh et al., 2006). Most of the literature in this area has focused on the implementation of these assessments and recommendations for instructional uses of benchmark testing results. For example, Supovitz and Klein (2003) recommend that benchmark test results can inform several stages of the instructional process including planning for instruction and teachers' decisions about content, pedagogical strategies or approaches and pace. Test results can also help to identify low-performing students and inform plans for additional supports or assistance to students who may be in need of remediation. More broadly, results of benchmark or periodic assessments can enable teachers to monitor and track student progress toward meeting instructional objectives and state content standards. In addition, results can be used to examine instructional effectiveness as well as the success of interventions or supplemental teaching used with underperforming students (Supovitz & Klein, 2003).

It may make most sense to think of benchmark or interim assessment as something that falls literally between yearly summative tests and daily classroom assessments, both in timing and in recommended use (Perie et al., 2009). It is a practice in which primarily summative tests are created and administered with the intent that results can be used formatively. Formative use of summative test data has always been a challenge, especially when formative assessment includes instructional correctives or other actions that are targeted to improve student learning. Perie et al. (2009) point out that to make instructional adjustments there needs to be close alignment with local curriculum and standards, an analysis of student misconceptions and misunderstandings, strategies for new instruction, and teacher use of the information to carry out the new instruction.

Research on teachers' perceptions confirms the purported utility of benchmark test results. When compared to state test data, greater percentages of teachers reported that benchmark test data were more helpful to "identify and correct gaps in their teaching" (Marsh et al., 2006, p. 5). According to Marsh, et al. (2006) teachers' favorable views were attributed to the frequent test administration, quick turnaround of results and the alignment of the benchmark test with the curriculum. However, when compared to their own classroom tests, 60 percent of teachers in one district reported their own assessments provided more useful information than the district benchmark test. They typically viewed their tests as more timely and thorough, and regarded the district assessment as taking away from instruction as well as providing redundant information (Marsh et al., 2006). Several studies indicate that teachers are using benchmark assessment results to make instructional adjustments, such as identifying and addressing areas of student weakness, providing remediation for gaps in student learning, setting instructional priorities and increasing efficiency, determining instructional approaches such as whole class instruction, differentiating instruction for small groups or customizing learning activities for individual students (Brunner et al., 2005; Christman et al., 2009; Marsh et al., 2006; Yeh, 2006). With regard to other impacts on teachers, the implementation of benchmark testing policies and expectations related to the use of the data have lead to reported increases in collaboration with colleagues and problem solving (Lachat & Smith, 2005; Wayman & Cho, 2009; Yeh, 2006). At the individual level both Brunner et al., (2005) and Yeh (2006) found teachers reported enhanced levels of self-efficacy and well as increased reflection on their practice associated with the implementation of benchmark or periodic testing systems. However, research suggests that the influence of benchmark testing is not consistent among teachers within a school. Marsh et al., (2006) found that some teachers use the information while others do not; 80 percent of the variability in teacher survey responses was within rather than between schools.

The central theory behind the implementation of benchmark testing is that the formative use of test results will inform instructional changes that will improve student learning. Empirical evidence that formative benchmark testing has a positive impact on student learning is both limited and mixed. For example, some research suggests that targeted instruction can lead to improvements in student test scores (Lachat & Smith, 2005; Nelson & Eddy, 2008; Trimble, Gay & Matthews, 2005; Yeh, 2006) as well as proficiency in reading and mathematics (Peterson, 2007). However, empirical investigations that utilized quasi-experimental approaches have found no significant differences between schools using benchmark assessments and comparison schools not using such tests (Henderson, Petrosino & Guckenburg, 2008; Niemi, Wang, Wang, Vallone, & Griffin, 2007). A descriptive study of 45 elementary teachers, using interviews, observations, and surveys, found that interim assessment data "did not substantially change their instructional and assessment practice" (Goertz et al., 2009, p. 6). They found that interim data did influence what was taught, but not how to re-teach. Other studies suggest benchmark testing can lead to positive impacts on factors that may ultimately contribute to improved student achievement such as increased student engagement and motivation (Yeh, 2006) and greater access to learning opportunities including tutorial and remediation instruction or services (Marsh et al., 2006).

In addition to the impact of benchmark testing policies on instruction and student outcomes, the literature suggests that a variety of factors are associated with teachers' formative use of benchmark test results, including the accessibility and perceived quality of the data. The timeliness and type of information teachers received was viewed as critical to the extent to which test results could be considered "actionable information". For example, online access to data was associated with teachers' use of data (Marsh et al., 2006). The RAND synthesis also suggested that teachers had concerns about the reliability and validity of test scores especially when teachers' perceived a lack of alignment of the tests with the curriculum as well as when they expressed concerns about students' trivial attitudes toward the test (Marsh et al., 2006). Other studies pointed to the need to provide capacity and professional development for teachers to support their use of benchmark testing data (Kerr et al., 2006, Murnane, Sharkey & Boudet, 2005; Symonds, 2004; Trimble, Gay & Matthews, 2005; Vogel, Rau, Baker & Ashby, 2006; Wohlstetter, Datnow & Park, 2008). These studies cited teachers' lack of expertise in analyzing and interpreting test score information and the need to develop a level of assessment literacy to support the effective and meaningful use of test score information.

There are limited empirical investigations of the impact of benchmark testing and more specifically the formative uses of benchmark test score results. The literature suggests that the way in which the results are used for formative purposes is highly localized, contextual and mediated by a variety of system- and school-level factors. Consequently, the present study was designed to examine teachers' formative use of benchmark testing data in several school districts surrounding an urban-metropolitan area in Virginia. The following research questions informed the study design:

1. How do teachers use the results of benchmark tests?
2. Do teachers' use benchmark testing data in formative ways?
3. What factors contribute to or detract from teachers' formative use of benchmark testing data?

Methods

To address the research questions a qualitative study was designed using focus groups as the primary method of data collection. A double-layer category focus group design was implemented, where individual focus groups were conducted with elementary and middle school teachers separately within a district, with the exception of one session that included a combination of elementary and middle school teachers (Krueger & Casey, 2009). A protocol was developed and piloted prior to the data collection. The questions in the protocol addressed four main topics: (1) the general nature of benchmark testing policies and the type of information teachers receive, (2) expectations for using benchmark test information, (3) instructional uses of benchmark test information and (4) general views on benchmark testing policies, practices, and procedures. Each focus group included 4-5 participants and lasted approximately 1-1.5 hours. The sessions were digitally recorded with participants' consent and the audio files were transcribed in preparation for data analysis.

A two-stage convenience sampling process was used to select and recruit focus group participants. Efforts were made to achieve maximum variation in the sample by identifying elementary and middle schools of varying Adequate Yearly Progress (AYP) performance levels from accessible urban and surrounding suburban school divisions. It was anticipated that including teachers from different performance-level contexts would yield focus groups sessions that described a full range of issues related to benchmark testing. Local school districts were contacted and informed about the study as well as the potential pool of schools identified by the researchers. District-level personnel communicated with school principals internally about their interest in the study and to seek permission to contact core-content area teachers for participation. Once individual schools were determined, a contact person within each school was identified who facilitated communication between the investigators and the teachers interested in participating in the study. Fifteen focus groups were conducted between the spring of 2009 and 2010 with a total of 67 teacher participants, representing six different school divisions in Virginia.

In Virginia the 2009-2010 AYP ratings were based on student achievement during the 2008-2009 school year. For a Virginia school or school division to have made AYP, at least 81 percent of students overall and students in all subgroups must have demonstrated proficiency in reading, and 79 percent of students overall and in all subgroups must have demonstrated proficiency in mathematics (http://www.doe.virginia.gov/statistics_reports/accreditation_ayp_reports/ayp/index.shtml). The state had overall pass rates of 89 percent in reading/language arts and 86 percent in mathematics for 2008-2009. With regard to student performance of participating schools, nine of the eleven elementary schools made AYP in 2009-2010. The average accreditation adjusted pass rate for English was 90.9 percent, with a range of 88-97 percent. Performance levels for mathematics were similar, ranging from 86-96 percent; the mean accreditation adjusted pass rate was 89.9 percent. Student achievement for participating elementary schools was at or above the state average; by comparison, student achievement for the participating middle schools was more variable and on average was below the state performance in reading and mathematics. At the middle school level, the average accreditation adjusted pass rate for English was 87.3 percent ranging from 81-96 percent; whereas, for mathematics, the mean was 83.3 percent with a range of 72-88 percent. Of four middle schools, three made AYP in 2009-2010. One non-traditional school was included in this study which also made AYP.

The majority of the participants were white (82%) and female (88%). While the participants were fairly homogenous with regard to gender and race/ethnicity they represented a varied degree of teaching experience. On average, study participants had been teaching for 11.5 years, with a range of 1-34 years of classroom experience. Roughly, one third of the participants were beginning teachers with 1-5 years of teaching experience while a smaller percentage (20%) had been teaching for over 20 years. The majority of participants were mid-career practitioners. In addition, two-thirds taught at the elementary level in grades 4 and 5 and the remaining third were middle school teachers in the areas of civics, science, mathematics and English language areas. All participants taught a grade and/or subject area that is tested as part of the state's Standards of Learning (SOL) mandated testing program associated with the *No Child Left Behind* test-based accountability requirements.

A transcript-based approach to data analysis using a constant-comparative analytic framework was used to identify emergent patterns or trends (Krueger & Casey, 2009). Throughout the data collection process the investigators met to discuss emergent patterns or trends, these preliminary themes were then explored in subsequent focus group sessions. After the completion of all of the focus group sessions, transcripts were analyzed with particular attention to the frequency (i.e., how often an idea was expressed) as well as the extensiveness (i.e., how many participants expressed the idea) of particular viewpoints or ideas using a coding system. Prior to analysis of the focus group transcripts, a priori start codes were developed and defined by the research team based on the related literature and the research questions (Miles & Huberman, 2002). Start codes were then applied through five rounds of trial coding and evolved in an iterative process. For each round, coders individually coded select transcripts and met as a coding team to collaboratively make revisions and additions to the code list and corresponding code dictionary to clarify and supplement the original start codes. Refinements to the coding structure were made based on group consensus. Thirty-four main codes were developed with a total of 107 associated sub-codes. Main codes were related to nine key categories: (1) Benchmark Testing Policy: length of testing program, test development process, alignment of test to SOLs, administration process, grading policy, (2) Receipt of Test Results: timing of receipt of results, level of reporting/analysis, type of data, analysis of data, format of results, who provides results, (3) Expectations for Teacher Use: district expectations, principal or building expectations, departmental or team use, (4) Instructional Uses of Results: identify students' strengths/weaknesses, identify curricular areas/topics to review or re-teach, determine approach to remediation of re-teaching, (5) Supports for Using Test Results: resources, specialists, principal, environment, accountability, (6) Obstacles/Barriers to Using Test Results: pacing, time, students, scoring error, test quality, technology, (7) Utility of Benchmark Testing: additional information provided, (8) Value of Benchmark Testing: overall value, impact on achievement and assessment, and (9) Recommendations: improve practice of formative use of test results, make results more useful/helpful. As an example, four sub-codes, *paper-and-pencil administration*, *online administration*, *quarterly administration* and *untimed testing*, were applied to capture the main code *administration processes* in these school districts. The *principal or building expectations for teacher use* main code consisted of six sub-codes - -

identification of weak students, identification of curricular areas to review, communication of test results to students, review of test with students, adjust instruction, re-teach content. One final example, *parent accountability, teacher accountability,* and *student accountability/motivation* made up the main code for *accountability* as a support for using test results.

Supported by ATLAS.ti (version 5.6) qualitative data analysis software, each coder used the revised code structure to code 10 percent of the focus group transcripts line-by-line independently to evaluate inter-coder reliability. With a high consistency in the application of the codes, each coder was assigned fifty percent of the remaining transcripts to code. Overall, 229 pages of transcriptions were coded for 14 of the 15 focus groups averaging 17 pages per transcript. Due to technical issues with audio taping, one focus group could not be transcribed. The coding team met weekly to review coding progress and to ensure the maintenance of common interpretations of the codes.

Results

Several themes emerged from the focus group data related to the general landscape of benchmark testing, expectations for teachers' use of test results, how teachers used test results to respond to student needs, factors that mitigated teachers' use of test results as well as systems or structures that contributed to teachers' perceived positive value of benchmark testing data. This section describes the themes and trends using exemplar quotations that reflect the majority view as well as minority viewpoints to present varying perspectives on benchmark testing.

Theme 1: Benchmark testing policies related to test construction and administration procedures were fairly similar among the school divisions. However, inconsistencies were evident across content areas and grade levels within school districts.

The vast majority of teachers described an online administration of the quarterly benchmark assessments. For those teachers in school divisions that administered the tests using a paper and pencil format, many indicated that the division was moving toward an online administration or administered the assessments online and in a paper-based format depending on the subject area. In addition, most of the benchmark assessments are developed "in-house" at the county level by a group consisting of instructional specialists, department chairs and/or lead teachers. It was not uncommon for the quarterly benchmark assessments to include SOL released test items or slightly modified versions of released items. Two teachers commented:

- *Our benchmarks are usually released items that have been slightly altered just so it's not exactly what they [students] would expect... but a lot of the times the graphics are the same.*
- *My [curriculum] specialist uses the same wording and the same examples as in the SOL tests. So it [test results] gives me a glimpse of how they [students] are going to perform on the SOLs. I am pretty happy about the way [they] designed the semester test.*

In other districts, teachers generated their own benchmark assessments by selecting items from a bank of test questions provided by the district. These test item banks typically included released SOL test items, items constructed by the instructional specialist as well as teacher-constructed questions. To illustrate, one teacher described this process:

- *We use a county-wide program which [curriculum] specialists have question banks…we use those banks to create our benchmark assessments, we test every three weeks, so every three weeks we'll give a test and those questions come from the exam bank.*

It was clear from the focus group sessions that there was a significant emphasis on not only aligning the content and format of the benchmark assessments with that of the state-mandated SOL tests but also on mirroring the administration procedures. For example, some teachers commented on the use of the state test blueprint to construct the benchmark assessments:

- *I usually find that the tests in Math are fairly reasonable. They [school division] base it on the percentages of the SOLs. The blueprint of the SOLs tells you that 10% will be calculation and this many [test items] will be that…the county follows that formula to make the test, so the test is not really surprising when you get it.*
- *It [benchmark test] tells you what you need to place more emphasis on. It really alerts you to the weaknesses of your class and how much more practice you need to provide. I think they are excellent, I think they are great because of their correlation with the SOL tests and this is one way of getting the children prepared and familiar with the formatting of the SOL test so that they will be successful.*

Other teachers expressed concern about the lack of alignment between the benchmark assessments and the state SOL test. For example, one teacher described the alignment of their division reading benchmark assessment as:

- *It's testing all the visualizing and all of the techniques that we teach, but that is not what's tested on the SOL…and so we're spending all this time worrying about* *what they get on this benchmark when the benchmark does not line up with the SOL test they will be taking in the spring.*

Similarly, a middle school teacher explained:

- *The SOL test for English is very different than the benchmark test. The kids are able to highlight, cross out things on the SOL...but on the benchmark they can't do that so when they are reading a whole passage, there is no way for them to delineate any information.*

This notion was echoed in several focus group sessions where teachers expressed concerns, particularly with reading passages, that students did not have the same access to tactile tools and strategies for the benchmark assessments as they did for the SOL test. These types of comments were not evident for other subject areas such as mathematics, science and social studies.

One area that emerged as a clear difference between elementary and middle school teachers was the use of benchmark test results in determining student grades. Teachers typically described the advantage of including benchmark test results in grades especially at the middle school level as a mechanism to increase levels of student motivation and accountability. One teacher remarked, *"I think that was the final argument in the middle school, was that if we want middle schoolers to study for it, they need to include it [the results in students' grades]."* In response to a question about barriers or obstacles to using test score information, a teacher noted *"...other than just kids not taking it as seriously as we'd like them to, that is probably the biggest barrier."* Some teachers discussed clearly established division-level policies while other suggested that there was some flexibility at the school level and teachers could decide how and if they wanted to incorporate student benchmark assessment results in quarterly grades. For example, one teacher explained:

- *They are graded, but they are not part of their grade. So they will [benchmark test results] show up on their report card as a separate category just so parents know and the students know what the grade is, but it doesn't have any effect on their class grade.*

Some grading policies were based on issues related to test security:

- *We haven't been able to use it [benchmark test] for a grade because we're not allowed to send them home to show parents so how can you justify that the child at the end of the unit was assessed at 80% when you can't share that with the parents.*

Alternatively, some comments indicated teachers' had flexibility to determine how the benchmark test results were incorporated into students' quarterly grades. One teacher explained:

- *A lot of that is actually left up to the teacher. It's not implemented county wide that it has to count as a test... it's been left up to the teacher [to determine] what parts of the tests they actually count...if you spent more time on one [content area] and didn't get to cover another, it is left up to the teacher to take those questions out, recalculate what their grade was based on what you taught.*

Theme 2: There are clear and consistent district- and building-level expectations for teachers' analysis and use of benchmark test results to make instructional adjustments in an effort to support student achievement.

Teachers described formal and informal systems that were associated with expectations of data analysis and use. Some described formal requirements to meet with principals or supervisors and submit action plans or reports to administrators that delineated specific strategies to support individual students' learning on the basis of test results. For example, a teacher described the requirement to complete an action plan as a plan for *"what you are going to do, how you are going to use the data to raise the students' skills and scores."* Other teachers' comments also describe a formal, specific expectation for how the results will be used:

- *We are asked to be accountable for each and every one of those students and sit face-to-face with an administrator and [he/she] says to you, how are you going to address those needs? And then we have to be able to say, well, I'm pulling them for remediation during this time, or I'm working with a small group or I've put them on additional enrichment, or whatever it is, but we've got to be able to explain how we're addressing those weaknesses.*
- *We have to use the results in order to figure out our weak students and what we need to teach them.*
- *They always want us to use it to just readjust our teaching.*
- *Well we have [name] meetings once a month. The department chair is in charge of that, basically, and then the central office. We discuss the data, we discuss the remediation plans that we need to do with the students.*

In contrast, other teachers discussed an expectation of informal processing of results:

- *Our principal expects when you have a grade level meeting to be able to say, this is what I'm doing about these results, because it is an unwritten expectation but it is clearly passed on, usually from mentor to mentee by sitting down with them the* *first time they are giving the test and describing how you do data analysis and literally walking them through it and showing them patterns to look for.*

- *I think it's expected that you, there's no question that you will use it and properly give all benchmark tests. And there's also an expectation of you to go back and do data analysis even though it's not a formal thing.*
- *It's up to the teacher how they're going to do it but the expectation is that you're going to get that child to pass the next text.*

Whether formal or informal, the expectations were clear and strong in schools where teachers indicated they used test results to re-teach and were generally positive about the test. The emphasis was on what they would do to remediate, to raise student achievement to result in higher end-of-year test scores. There was relatively little expectation that student learning would be enhanced beyond what was perceived as needed for successful accountability testing. This was reinforced by how most teachers indicated a use of results that focused on students who showed poor performance – particularly those that are on the borderline of passing, not for students who did well:

- *My biggest problem I have with it [benchmark testing] is that we look at these kids – these kids are doing good so we're not going to really worry about them. These kids are really, really low so we're going to just kind of ignore them, and we're going to take these kids that are on the borderline and these are the ones we need to work with to get them to where they can pass. I feel like the ones that are really, really low are not getting enough help...you look around and you go, well, you know, this kid's really, we call them bubble kids, that group in the middle that we are really focused on, and I feel like the kids that are really, really* *low are just getting left because we are focusing so much on those ones that we want to push over that mark.*

- *We look at the failures and determine what we need to work on, what we need to do some re-teaching on.*
- *They [test results] are nice because it can help me see, like everybody is saying, something that I need to review. If I haven't like hit geometry ... it will help me see that.*

Theme 3: Timely access to test results and use of a software program supported data analysis and reporting.

The vast majority of teachers had almost immediate access to their students' benchmark test results through the use of a data storage system or software program – *"the fact that you can get your results back so quickly is very helpful…and that is a real positive."* Teachers also described these systems as user-friendly and flexible. They could access results on different levels, such as examining how their class performed as a whole by conducting item analyses and then reviewing the performance of individual students. Few teachers mentioned that they also disaggregated results by different sub-groups. The most common ways teachers disaggregated results was by SOL strand (i.e., content standards). One teacher explained:

- *We have a scanner here, and then we log in and, it's like item analysis, when we put the test in, we put in the SOL or whatever the strand is and then we'll look at the results for strengths and weaknesses, and if there is a weakness it may be highlighted.*

Similarly, others commented on the value of a computerized data system, not only for the ease of obtaining reports and options for reporting different scores, but also the ability to obtain immediate results.

- *You can break it [benchmark test results] down by SOL strand as well on this computer system. And I think it's important because it can show you whether or not the class as a whole did well or if it was just that one particular student where you need to go back and remediate, or it could reflect on you as a teacher, maybe I didn't teach that particular concept thoroughly enough and it might not necessarily be the child which I think is really helpful… It helps you know what you need to hit on for the SOLs.*
- *That if we are supposed to be using this information to guide instruction we need immediate feedback, like the day of, so we can plan to adjust instruction for the following day.*
- *You can do average score, average score by class, individual students like exactly which ones [test items] they missed, what letter they put so that you can kind of talk to them, why they put that letter.*

In addition to describing the utility of the different computer programs, a few teachers expressed frustration with the time required to analyze student test results – *"…filling out the spreadsheet doesn't take very long, it takes long to sit there with your results and decide what to do."* In addition, a perceived lack of guidance or direction at the district level was also a concern among teachers. The following comments illustrate this view:

- *I think part of the frustration we feel is we're not quite sure what the county is trying to do with it [the data] either. I don't know what their vision of it is. We* *look at the data a lot, we go to [subject] meetings and they're like, here's the data. We had a professional development day and a [subject] meeting two days apart, and a faculty meeting and we talked about the data at all three meetings but I walked away from those saying, I really didn't get anything out of this.*

- *In a faculty meeting, they [school administration] will present all of this data, and the entire faculty is sitting there looking at math and language, and science and social studies, and after an hour they will say okay, great. Okay, tomorrow break down in your departments and do the same thing again; then break down [the data] by your grade level. I think part of the problem is that they aren't exactly sure what they want to do with it, so therefore they are not answering, they don't have questions that they wanted answered.*

In addition to having immediate access to students' test results and a variety of tools or options for data analysis to make appropriate instructional adjustments, there was also a consistent expectation for teachers to review the tests and provide immediate feedback to students:

- *We are required to go over each test with our students. I know a lot of [teachers] that immediately go over the test with their students, their children mark on their test booklet as well as their bubble sheets, so as soon as they turn in their bubble sheets, they [teachers] go over the test with them, so they can have the test in front of them. A lot of times I'll do that before I get the results back because you can immediately get them feedback.*

Teachers also expressed frustration with the time required to review benchmark tests often at the expense of moving forward in their instruction and staying on track with the pacing or curriculum guides:

- *Some of the [questions] take a while to go through and have them look back at the passages, so you are talking two to three days lost, two or three periods of time lost just for one test.*
- *I've got to say we spend way too much time looking at data, I mean, and not enough time actually doing something with it or about it. It's like data, data, data, meeting, meeting, meeting, okay, go teach again.*

Theme 4: It was important for teachers to discuss results with others and have time with colleagues to discuss results.

An essential part of teachers' analysis and use of the benchmark test results was the opportunity to review and discuss the results with colleagues, often informally. Rather than analyze and interpret results individually, teachers found that a team approach or informal meetings with teachers who have the same class was important. For example, one teacher commented:

- *We have achievement team meetings where we look at every single teacher, every single class, everything, and look at the data really in depth to try to figure out what's going on. What is the problem with this class? Why is this one doing better?*

Teachers described the support system that is evident in meetings and how the discussions led to appropriate interpretations. Comparing results across teachers was key to this process, as well as opportunities for review that were not judgmental about their teaching effectiveness. As long as there was an emphasis on whether incorrect answers meant that students did not understand or know something followed by discussion of next steps, teachers found the tests helpful:

- *I think we just review what it is [the results] and we say, hey, your kids did really good on that. What did you do to make them understand? Maybe mine didn't do too well. We exchange ideas, and brainstorm together. We work as a team.*
- *I have taught at many different schools and I don't know if this is unique to [this school] but you don't have this shame issue of okay I have six that aren't meeting my standards. It is very open and free, I've got these six [students] that aren't meeting the [standards], this is what is going on, sometimes they are performing, sometimes they aren't...there isn't this [culture] were we have to hide this, we really seem to be a community that is pushing together to help our kids achieve.*
- *We usually split into sections you know the math people will talk about the math stuff, and split off and come back together and we talk about ideas of what can help, what kind of remediation we can do.*

Teachers' described these meetings as being informal as well as systematic and scheduled on a regular basis. A few teachers communicated that they conducted their data analysis independent of their colleagues, as illustrated by one teacher: "*we don't do anything with it [test results]. To be very honest with you, we as individuals look at it [test results], we look at the item analysis, but as a department we never come together.*"

Theme 5: Many teachers claim to analyze benchmark test results at the class and individual student level to inform review, re-teaching, and remediation or enrichment. Individual student versus class needs guide teachers' next steps.

Many of the teachers maintained that the test results identified weaknesses in student learning and subsequently led to re-teaching and remediation. There was an emphasis on identifying gaps in learning once the results were confirmed to indicate a lack of student knowledge and understanding and not due to poor questioning. It was clear that the results did not directly imply lack of knowledge or the need for re-teaching. Rather, results were appropriately used as a *possible* indicator of student or instructional weakness, and only on further discussion did teachers get to the point that actual re-teaching was implemented:

- *I'm looking to see what they mastered and where the weaknesses are per individual and per class.*
- *I definitely look at how many kids missed a certain question. Did the majority of the class miss this question? Can I see a trend with this topic? I don't want to say that I ever brush any skills off to the side, but I do hit those weaknesses and the gaps hard.*
- *And they [administrators] would look at the failures and determine what we need to work on, what we need to do some re-teaching on. But once the data, once we get the data back, usually we put in the lesson plan to over it with the students, re-teach, and use strategies that they may have used to help them.*
- *The seventh grade math [team] has to sit down and figure out the five most missed, and if we're going to put them in our warm ups, or we're going to put*

then in our bi-weekly set of quizzes. We definitely do an awful lot of data analysis and change our instruction.

- *It makes a difference in my instruction. I mean, I think I'm able to help students more that are having difficulty based on it. I am able to hone in on exactly where the problem is. I don't have to fish around.*

Class-level weaknesses are often addressed through review of the benchmark tests as a group concentrating on missed items:

- *It can show you whether of not the class as a whole did well or if it was just that one particular student where you need to go back and remediate, or it could reflect on you as a teacher like maybe I didn't teach that particular concept thoroughly enough.*
- *We go over the benchmark tests after we've graded them... If it's one that a lot of people missed, [then] it would be one that as a class I would go over...If it's just one or two people that missed it, that would be one you would pull them individually and work on.*
- *If I see a large number of my students missing in this area, I'm going to try to re-teach it to the whole class using a different method, or a different learning style or something. If it's only a couple of ones, no, I'll pull the aside and then instruct one-on-one.*

Teachers incorporate these items into homework and quizzes to provide additional practice targeted to individual students as well as the whole group –

- *If it tends to be several students that missed the [questions], I have these kids do three review items every day. If I feel like there's a bigger chunk in the* *classroom, I use it for my review questions and then put them in some homework just to get them to see it more and more.*

- *At the end of the first 9 weeks, I saw what everybody's weakness was, and I put the extra homework in each, in their individual folders and that way, what I did not get them to cover for the first 9 weeks.*

One mathematics teacher created new practice items from missed benchmark questions:

- *That tends to drive my warm-ups, the questions that they missed. It's the same [question], but I might change the numbers around, and that tends to help them out.*

In addition to review of content based on test results, teachers also use results to guide their review of test taking strategies and multiple choice question formatting to prepare students for subsequent benchmark and SOL tests:

- *The one thing is showed me…is the kids who were struggling with tests, and to me, art of it is more like testing and strategies for kids who are weak readers, so it flags kids that I need to start pulling aside.*
- *We can go into a lot of test taking techniques. I think we really have a lot. I post it in my room, and we stop and talk about those, what you should do to approach this test.*
- *We try to give them extra practice, we'll practice on that particular formatting of the question so that when they see it again, they'll be ready for it.*
- *She [instructional specialist] may give us more [questions] in the next lesson plan, more of what the children should have more experience with, those types of questions.*

Teachers used the results to understand student performance and learning needs and subsequently provide appropriate re-teaching or remediation in response to these needs. As teachers noted:

- *Once we get the data back, we put in the lesson plan to go over areas we need to work on with the students, re-teach, and use strategies that we may not have used to help [students].*
- *When we design lesson plans...I would usually look at the scores and divide them into groups, who needs the prerequisite, who needs to go back and review, who is meeting the initial requirement, and who needs to move on to other things. That grouping basically determines how the remediation is going to move on.*

One teacher noted the value of using benchmark test results as an indicator of what content students learned and retained:

- *I can see who's not retaining information. Maybe they understood it long enough to regurgitate it for my [classroom] assessment but really don't have a strong foundation to remember it later down the road...[if] they're not showing it [skill] on the benchmark then I can go back and use that to harp on that skill some more.*

A number of schools also support structured instructional time or tutoring for remediation and enrichment based on individual student needs as demonstrated by performance on benchmark tests. These types of support programs often occur during the school day whereas most tutoring activities occurred after school. Comments made in the focus groups indicate that teachers and administrators use results of the benchmark tests to identify students and make appropriate placements. The following comments illustrate how teachers make use of available programmatic resources based on information provided by test results:

- *I use the results to pick out children to stay after school with me for after-school tutoring and small group separate instruction.*
- *I think it's been very helpful particularly in identifying which kids need extra tutoring.*

In addition, teachers described selection of content and instructional strategy for group-based remediation and enrichment informed by test results. For example:

- *We use that data from the test to decide what people should be doing when they're part of the [name] learning center. We'll use data to give to our tutors in the afternoon as well so they'll have specific strands the students can focus on.*
- *We have a block [for enrichment] every day for 45 minutes called [name]... We have four different groups since we have four different teachers...We look at what as a whole our kids struggled with, if they did [struggle], and we group the kids...and do a lot of project based activities. The kids that are doing really well, they do an enrichment type project.*

Results from benchmark tests also serve as a means for reflective practice for teachers, enabling them to evaluate their own instruction. It was clear from talking with teachers that they used the benchmark test results for multiple purposes to make assessments about student progress as well as to examine the efficacy of their teaching. The following comment reflects this concept:

- *You can break [results] down by SOL strand on the computer system. I think it's important because it can show you whether or not the class as a whole did well or if it was just that one particular student where you need to go back and remediate, or it could reflect on you as a teacher. Maybe I didn't teach that particular* *concept thoroughly enough…If the math test was multiplication, you can know, I didn't teach that very much so it helps you know what you need to hit on for the SOLs.*

Review of performance on particular SOLs also encouraged teacher self-assessment, for example:

- *If I see that my students have scored low in a particular SOL strand, then I know that's an area that I have not addressed with as much strength as I should have.*
- *If we're missing a certain SOL, we have to decide is it something as a teacher that I didn't cover or that the kids didn't quite get or that certain individual students are missing. From there, we have to spiral back and make sure they all understand the information.*

Another teacher elaborated on the use of test results to inform not only current instruction but subsequent instruction in the upcoming academic year:

- *We also look at our own test data from the previous year to see if there was an area of deficit for ourselves as a teacher. Did I not teach matter well enough? Did I not teach long division well; what strands were my children overall missing and what can I do to improve that as well? So, the year starts out with data analysis.*

Although most teachers make use of benchmark test results to inform practice, evaluate student learning, and self-assess, some teachers described struggling to find time to effectively use results to adjust instruction. One teacher is challenged to find time to review the benchmark results thoroughly with her students:

- *[Teachers] are supposed to give the benchmark, go over it with the kids, and get them ready for their SOL…It's too much. So, how much are we really getting from the test results? How are you using that in your class? From my perspective, I'll pick out some questions for my classes to go over, but I just can't give up a day or a whole block or two to go over it piece by piece.*

Two teachers described the difficulty of freeing up time to remediate as they aimed to stay on schedule with their pacing guide:

- *The results give you a snapshot of where they are at that point in time…but in following the pacing guide that has been suggested, I don't have time to go over it until right before SOLs. So, it does give me somewhat of an idea of areas where [students] are having problems, but I'm not going to be able to really truly address that until the end of the school year when I'd be reviewing it regardless.*
- *The problem is…by assessing and spending time on that, you lose the time you could actually remediate because you need to move on…we need to redo certain content because the benchmark data is showing us that [students] didn't get it, but where does that time come from?*

One teacher found more general frustration with unclear expectations of how to utilize the results effectively given existing time constraints:

- *If we had a better idea of what to do with the information from a benchmark test, for example, after you give this test you will have days built in to go over it or re-teach. But, if you're just giving tests as you go through, there isn't time to do anything with it, and then it's like you're giving a test for sake of giving a test.*

While there were over 100 comments from teachers about the use of the results to remediate and re-teach, there were also several teachers who indicated a negative view of the usefulness of the results, the emphasis on multiple choice items, and more broadly the entire program. Elementary and middle school teachers expressed discontent; however there was greater intensity among middle school teachers. Some participants indicated that the amount of time devoted to benchmark testing is excessive in relation to what is learned. In one case a total of 12 instructional days during the year was devoted to benchmark testing. In several of the focus groups teachers were asked if they would be upset if benchmark testing was eliminated. Virtually all responses were negative – they would not be upset. For some teachers it would be "great" if benchmark testing was eliminated; for others it seemed to be "okay" or even "good" to keep benchmark testing since it did, occasionally, help identify student weaknesses, and this would hopefully lead to higher end-of-year test results. Some teacher statements that represent negative opinions about benchmark testing include the following:

- *It does get overwhelming to have to continually prepare the tests ... it can be a little much.*
- *I don't like that it [benchmark testing] takes away from the creativity of other things that I would like to be doing ...because it's multiple choice and you've got to fit everybody into this little box and get those little multiple choice answers done and it's a struggle for me because there are other ways I think kids can be assessed.*
- *We aren't able to construct the test ourselves ... I found out that there's too much, there are too many holes in this benchmark test.*
- *Let's just eradicate benchmarks all together. That's my opinion.*

- *Take them, bye bye.*
- *Yeah, take them, it's one less headache we have to jump hoops through.*
- *There's got to be a better tool of assessment than a benchmark because to me it doesn't hold a whole lot of value.*
- *It's not teaching them life skills to teach them how to take a multiple choice test. That's not preparing them for life...a multiple choice test in my opinion does not provide quality information for us to be able to move forward with these kids.*
- *I just don't like it. I want something that's going to help me out at the end of the year.*
- *I hate to say it but I don't think the benchmark really tells me more than I already know before they took it ... basically for us I think the benchmarks more or less confirm what we already knew.*
- *For me, it's not worth the amount of resources that we're dedicating to it.*

Theme 6: A variety of factors impact teachers' use of benchmark test data, including the alignment of the test with the content of instruction, quality of the test items, the accuracy of the scoring, and the technology available to support the test administration.

The vast majority of teachers in the focus groups described having a pacing or curriculum guide that identified the specific content standards that should be taught each quarter. The benchmark tests are designed to assess this content. Teachers' views on the extent to which the benchmark tests were aligned with the pacing guides were mixed; some expressed positive opinions while others were concerned about the lack of alignment. For example, when asked about the issue of alignment, one teacher explained, *"they [benchmark tests] line up with our* *pacing guide that we are supposed to be teaching each quarter."* Another group of teachers responded that the alignment between the tests and the content was unanimously *"right on"*. They further described the alignment with both instructional content and skills, *"that you have to teach [the content] within that nine weeks because it goes with the benchmark...and especially with the strategies, the keys to comprehension...visualizing and [making] inferences."* Alternatively, some teachers expressed frustration with a lack of alignment:

- *The other problem too is when you have your pacing guide and they tell you to hit this [content] the first nine weeks, a lot of times the questions on the benchmark aren't correlated with what you were teaching the first nine weeks, so they will have questions about things that they didn't tell you to go over.*
- *I think the tests are pretty fair, but sometimes they don't match up with our pacing [guide] and that makes it unfair to the child.*
- *We had one assessment [benchmark] that according to the pacing guide we were supposed to have completed, the division of decimals or the multiplication of two digits...and the assessment had nothing on it. So they should line up with what the expectation is for us for teaching, and they don't always line up.*

Another teacher described alignment concerns related to the nature of the end-of-year SOL test:

- *The other issue in Science – 8th grade physical science – is that the end of the year SOLs for 6th, 7th, and 8th grade are combined into [this one test], well the third 9 week benchmark exam which was typically a pre-SOL practice test, turned into a real graded [test]. The 6th, 7th, and 8th grade material on it was very limited on what we had taught in that third 9 weeks for physical science so it was like your students are being asked questions on 6th, 7th, and 8th grade material but very little* *on what you just actually covered, yet we are calling it a third 9 weeks assessment. So it was a real problem.*

Several teachers noted that over time, as the benchmark testing program had been implemented the alignment had improved. For example:

- *I will say this though, that our pacing guide since I've been in the county has probably been more aligned with the benchmark…when we are told to give the benchmark, we have at least covered the SOLs, where there was a time that was not so.*
- *They [benchmark tests] have gotten much better in the alignment because I remember it was only my second or third year [teaching] I was asking what's going on, am I not doing what I should be doing? But it wasn't just me, thank goodness.*

The negative views on the alignment tended to be associated with recent implementation of benchmark testing, such as programs that had been administered for only one or two years. Teachers' perceptions of the quality or integrity of the test was also a factor that influenced the extent to which test results informed instructional decisions. Many teachers commented on having dealt with scoring errors, poor wording of questions and a lack of content representativeness. For example:

- *Last year's benchmarks…were just full of errors and questions that didn't cover actual grade level material…I don't really think it [the benchmark test] was good, probably wasn't very useful.*
- *Sometimes those questions that are in there are just simply bad questions.*

- *In general it's not because it's the information, it's because of the way the question was worded.*
- *We really need to focus on the tests being valid. It is hard to take it seriously when you don't feel like it is valid. When you look at it and you see mistakes or passages you know your students aren't going to be able to read because it is way above their reading level. The people writing the tests need some kind of training.*
- *Sometimes the terminology isn't maybe appropriate for that grade level, and it doesn't fit with what you typically would teach that grade level of kids.*

If large numbers of students missed the same question either in the same class or at the same grade level, this was often a red flag or signal to teachers that the wording of the question was problematic, a scoring error had occurred or students were tested on information that they hadn't had the opportunity to learn. For example:

- *In general if one class misses it [test question], generally they all miss it. And it isn't because it is the information it's because of the way the question was worded.*
- *Sometimes, I think ...you have to use your professional judgment, these are not norm referenced tests and sometimes the questions that are on the test are just simply bad questions. I think there is a certain amount of professional judgment if a child [doesn't understand] cause and effect, was it because they didn't understand cause and effect or is it because that was really a poorly written question?*

Teachers also tended to question the validity of the conclusions they could make about students' levels of content knowledge and skills if the test did not contain a sufficient number of questions in certain areas. The following statement illustrates this view:

- *Many times the 9-week assessments are so all encompassing that it is difficult for the students....you may only have one question that addresses a specific objective and so that is not really a true representation of what the child knows about that objective.*

Another issue that teachers indentified as influencing their use of test results was technological obstacles. Teachers talked about technology in several different ways, including issues related to their use of software programs or databases to analyze test results, "glitches" in test administration that comprised the validity of students' test scores, test formatting required for online administration and general access to computers. With regard to teachers' capacity to effectively and efficiently analyze test results housed in an electronic database system they indicated:

- *While we were all given [training], the training was minimal...Some people are good at going in and learning a program on their own, but if that's not where they're comfortable, I think some people chose not to do that. And because it [use of the database] wasn't mandatory, this is how you get this [type of analysis], a lot of people have never figured out how to do that. I think that's one area that we're possibly lacking is training and how to get specific kinds of information.*

Alternatively, another teacher described the relative ease of using the data base program albeit the process seemed unnecessarily complicated,

- *I think it is easy to understand because it's kind of limited in how much you can do. But to get to one test, you have to click at least six different places, you click this, and you go to this, so to explain to someone is also difficult. Once you've done it you know [after] I think the first few times, I just accidentally found what I needed four different ways.*

Technological obstacles were also described in the context of the online administration and format. One teacher described the apparent impact of a storm on her students' online testing experience this way:

- *It was taking 15 minutes for the next question to appear and then the answers wouldn't be there and so we just stopped. But then because they had started [the test], that was recorded and we could not go back in and do it again. So it showed for my whole class, my average was a 20 or something like that.*

These types of comments were less frequent than those related to the online formatting of the test which typically focused on reading passages:

- *Reading passages for English, we gave a hard copy simply because we were having a little bit of trouble with [online administration program] so we gave them the large passages to read in hard copy, but the questions themselves and answer choices were online. For English, we need to split the screen a lot and read the passages and that presented a problem.*
- *Having a child read a passage that's five or six paragraphs long sometimes is something that all of our kids aren't prepared for or ready for and when it gets time for SOLs they don't have the paper copies that we give them. They're just reading through what's on the computer screen the entire time.*

- *I think the online test is bad for some students especially in reading the passages because they can't just keep up and go back and flip up or use any of their reading strategies and those grades are normally lower when they do it online as opposed to using a hard copy.*

A teacher described the preferences of her students with regard to the online test format:

- *A number of my seventh graders, their feedback that I received in class and in chatting with them after the first time they took this last benchmark was, [name], we much prefer paper and pencil. I asked them why, and a lot of the same things we are hearing came out, they like to highlight, what they see on the screen and what they down on paper might not match.*

Other teachers expressed similar limitations of the online benchmark test compared with the tools available when students take the end-of-year state SOL exam – "*they will be able to highlight on the SOL*" and "*…they did not have those tools last quarter on the computer and that was a big challenge for these kids because they are so used to it.*" Other comments related to access focused on the limited number of computers available to administer the online benchmark test, for example:

- *If we had one computer for every student that was testing at that time, that would cut down on a lot of the lost time because we could tweak our schedule a little better, but because we are limited with the number of computers, we do a lot of flip-flopping with third, fourth, and fifth [grades], …that's where a lot of our time is lost, is having to wait for somebody else to finish a test…if every kid had a computer to work from, everybody could test at the same time.*

Theme 7: Teachers expressed significant concerns about the amount of instructional time that is devoted to testing and the implications for the quality of their instruction.

Although most teachers described aspects of the benchmark tests as useful to identify students' strengths and weaknesses as well as the limitations of their own teaching, they voiced significant concerns and frustration about the amount of time devoted to testing and the loss of instructional time as a result. For example:

- *Just the time is takes to give all these assessments. As important as these assessments are, it does take instructional time... we don't just do those [benchmark assessments], because we do a lot of pre and post-assessments so this is just one more thing on top of a lot of other testing we do.*

Other teachers expressed concerns about the impact on teaching effectiveness:

- *I think it is definitely made us change the way we teach because you are looking for how can I teach this the most effectively and the fastest...that is the truth, you have got to hurry up and get through it [curriculum] so that you can get to the next thing so that they get everything [before the test]. I do feel like sometimes I don't teach things as well as I used to because of the time constraints.*
- *You are sacrificing learning time for testing time...we leave very little time to actually teaching. These kids are losing four weeks out of the year of instructional time.*
- *It's at least a 3-day process in terms of the testing, and I just haven't seen information that's of that value when you consider that you get 90 days with each of your classes and if you are spending 3 days each 9 weeks, 12 days out of the*

year, that's an awful lot of time that you're spending trying to develop information for which I am getting minimal benefit.

Teachers described feeling pressure to cover content, and at times the seemingly overwhelming about of content they had to teach. A teacher described:

- *I would think it might be possible to maybe eliminate one of the reading or one of the math benchmarks, just eliminate one more week where you lose instruction because every minute I have to teach my kids I like to have to teach my kids.*

Regardless of the constraints on instruction in an effort to accommodate benchmark testing, teachers recognized the potential benefits of the information. For example:

- *It [test results] helps me analyze my instruction as a whole. I didn't present the questions in my classroom the way the questions are presented on this test. So maybe I need to back track or maybe I didn't spend a lot of time in this area, and I need to go back and [re-teach].*
- *They are one more piece you can have and when you compare it across the county you can use it to see how you are doing...but again, it just shows that it isn't the only thing we base our instructional decisions on by all means.*

Discussion

The results of this qualitative inquiry of teachers' use of benchmark test results to support instruction and learning are, in the main, consistent with those of previous research on this topic. Teachers described using benchmark test results primarily to identify students' strengths and weaknesses and for determining whether additional instruction was needed. Additional instruction usually involved review or re-teaching for the whole class or developing small group or individualized programs for more extensive remediation depending on the extent of student misunderstandings. These findings are consistent with those of other similar investigations (Christman et al., 2009; Brunner et al., 2005; Marsh et al., 2006; Yeh, 2006). Teachers also described using benchmark test results to examine their own instructional effectiveness and to identify areas in the curriculum that may require more intensive instruction or the use of different pedagogical strategies. These results illustrate the practical applications of the recommendations provided by Christman et al. (2009) and Supovitz and Klein (2003) regarding the potential impact of benchmark tests on teachers' practice.

Teachers in this study detailed the benefits of having opportunities to hold conversations with their colleagues about benchmark test results, and engage in critical analysis and interpretation of student test scores. Similar to the findings of Christman et al. (2009), Lachat and Smith (2005), Wayman and Cho (2009) and Yeh (2006), these conversations often led to collaborations that involved sharing successful instructional approaches to address common areas of weakness and developing specific plans for providing remediation either during or after school to individual or small groups of students. Collaborations seemed most successful and meaningful when there was a supportive culture or school environment in which there were clear expectations that teachers discuss results and consider what instructional activities would be helpful in closing the gap between current and targeted understanding. Teachers working in schools where there was considerable pressure to raise student test scores were less likely to describe these conversations as meaningful or as contributing to their own professional growth. Discussions about benchmark test data were described as most beneficial when meetings were held regularly and structured time was incorporated into the school day to facilitate professional conversations.

Similar to the findings of Goertz et al. (2009), although many teachers described the review or re-teaching of content associated with student weaknesses, few described specific instructional changes in detail. Often, teachers would describe providing additional homework, weaving short reviews in the form of questions into warm-up activities, or providing workbook exercises or worksheets, but there was little discussion of specific modifications to the delivery of content or instructional strategies. The extent to which the benchmark test results are having more than a surface-level impact on instruction is unclear. The degree to which the test information influences instruction is most likely due to teachers' perceptions of the meaningfulness of the information once interpreted in light of the quality of the items and discussions with other teachers to identify trends. We found, consistent with recommendations from Christman et al. (2009), that teachers could benefit by professional development to know how to frame conversations and questions about assessment data so that appropriate implications are identified and inappropriate uses are avoided (e.g., predicting end-of-year results). For example, do discussions deepen teachers' understanding of student progress and their own instructional practices? It is possible that providing some structure to help facilitate teacher discussions would be helpful (e.g., in the form of a list of questions and reminders about invalid uses; a list of suggested remedial with strategies with strengths and weaknesses of each alignment between the nature of the misunderstanding and instructional alternatives for "re-teaching").

There is clearly some question about whether benchmark testing provides sufficient benefits in light of the resources that are needed and the instructional time that is supplanted by more testing. In Virginia state accountability tests are untimed, and usually the benchmark testing would mimic this condition. Some students are allowed to take a full day to complete a

50 item multiple-choice test. As a result, in some schools many days would be used for testing. This may not be such an issue in other states where timed tests are used. Nevertheless, it is appropriate to ask questions about relative benefits of using time dedicated to instruction for benchmark testing or other ways of evaluating student learning.

Our findings suggest that administering benchmark tests online, with immediate reports that teachers could format, is an essential though an insufficient element of positive use of data. The immediacy of receiving reports seems to be important, especially for reviewing test results with students. We suspect that delayed reporting results in less effective use, though our data do not specifically address this issue. With increasing sophistication in software, computer availability, and teacher skills in working with test results, a corresponding increase in effective use of benchmark test results will occur.

The reliability and validity of the test score data were significant factors in the extent to which teachers' used the test score information to make instructional adjustments or to develop individualized plans for students. The quality of test items, scoring errors, and a perceived indiscriminate implementation of benchmark tests substantially undermined teachers' perceptions of the usefulness of the test scores and the influence brought to bear on their instructional practice. It was essential for teachers to view the items when reviewing results. This was needed to assess the quality of the item and alignment with teaching, and how students were asked previously about their understanding. In this type of context, it was not uncommon for teachers to also express concerns about the alignment of the benchmark test with their classroom instruction and/or the content or types of items found on the end-of-year SOL state exam. As pointed out by Christman et al. (2009), however, it is possible that too much emphasis on individual items leads to an inordinate amount of time spent on test questions, with less emphasis on understanding student competence and designing further instruction to close learning gaps. Furthermore, when coupled with the extensive amount of time required for the test administration, data analysis, and often times the required review of test results, the frustration among teachers was significant. Teachers who questioned the capacity of the tests to provide new information about their students or perceived the results as suspect or flawed, often expressed negative views about the usefulness of the benchmark testing program.

While the majority of teachers described the information provided by benchmark tests as helpful in identifying weak curricular areas, the extent to which these data were used in formative ways is less clear. It is understandable that when viewed as not credible, test scores would have less influence on instruction than information resulting from high-quality assessments well-aligned with the instructional content. Even in the most desirable situations it was unclear the extent to which benchmark test results were serving truly formative functions as defined by Wiliam and Leahy (2007) and considered by Popham (2008). Wiliam and Leahy emphasize that "formative assessment" is not formative unless the information can be used to improve the quality of learning. Teachers often questioned and were uncomfortable in many cases about the impact of testing on the quality of their instruction. As noted previously, a teacher described the impact on instruction:

> *I think it has definitely made us change the way we teach because you are looking for how can I teach this the most effectively and the fastest…that is the truth, you have got to hurry up and get through it [curriculum] so that you can get to the next thing so that they get everything [before the test]. I do feel like sometimes I don't teach things as well as I used to because of the time constraints.*

The extent to which teachers have the opportunity to use the results to improve instruction is unclear.

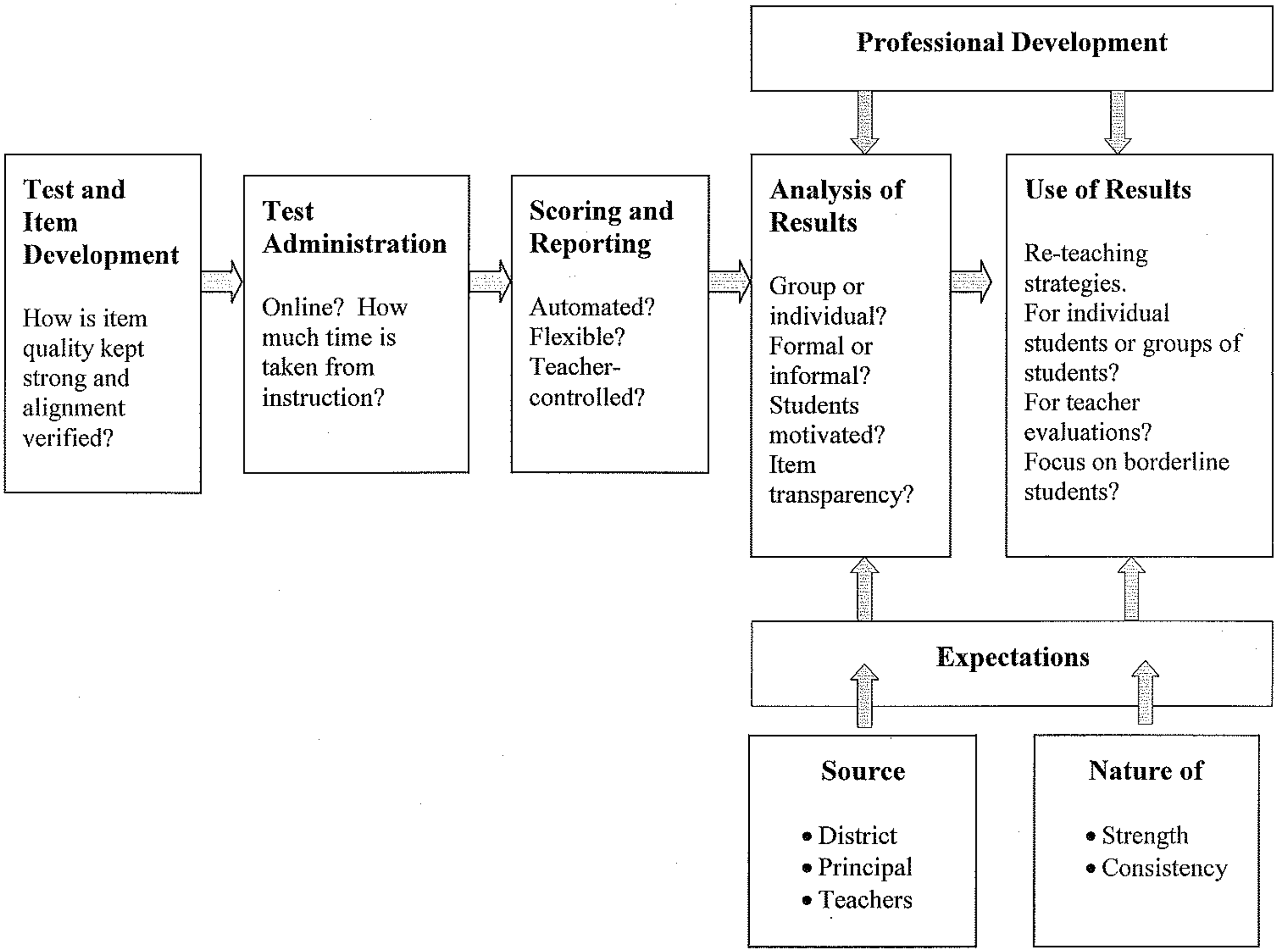
Popham (2008) suggests that formative uses of benchmark test data require teachers to modify and "adjust their on-going instructional procedures" (p. 6). Teachers described varying degrees of instructional responses to benchmark test results, most of which were mitigated by perceptions of data quality, however very few detailed making marked changes to their instruction on the basis of test results. While the teachers in this study were profoundly passionate about supporting their students' learning, and many were regarded as instructional leaders in their grade level or content area, there is only moderate evidence to suggest that they are using results in ways that would be described as formative as defined herein. While there was clear potential for using results in formative ways, the most significant constraint for doing so seemed to be the lack of time – time to thoughtfully analyze data, time to meet and collaborate with colleagues, and time to provide high-quality instructional correctives. Teachers' comments suggest that the demand to cover content and remediate low-performing students while under the pressures of decreasing time for instruction was a significant and almost insurmountable challenge. As a result, it is not surprising that teachers did not describe instructional adjustments in greater detail or specificity, although the potential and capacity for making formative adjustments were clearly present.

In an effort to systematically combine findings from the themes, key components of using benchmark tests in formative ways are illustrated in Figure 1. This figure shows the sequence of essential factors that appear to influence the extent to which benchmark tests are used effectively for "formative" assessment (clearly benchmark testing and subsequent use of results for further instruction is not consistent with formative assessment as a process that occurs during teaching).

The figure shows some operational factors that seem necessary, but not sufficient, for successful use of benchmark tests, such as high quality test items, the use of online, automated test delivery and administration, scoring, and reporting, and other factors that could vary significantly even if appropriate technical aspects are in place. Also illustrated are teacher descriptions of how test results were used in their schools and districts, which included both intended and emerging uses of benchmark test scores. One emerging or perhaps unintended use was evaluating teachers on the basis of student test results. In districts where this was occurring, teachers expressed significant frustration, often times commenting about having received inconsistent and contradictory messages about the purposes of the benchmark testing program. For teachers who expressed concerns about the quality of the benchmark assessments the use of test scores for evaluation purposes was viewed as highly problematic. Clear and consistent communication about the intended purpose and use of benchmark test scores is fundamental to effective teacher use. If test scores are being used to make decisions about teaching effectiveness, which is arguably a questionable practice, the use of results in this way should be an intended purpose at the design and development stage of the tests. Professional development is identified as a critical input. With the essential need to include informal teacher discussions about results, the focus of this professional development needs to be on what influences student performance, how test scores are interpreted, how levels of performance are supported with other evidence, and, perhaps most important, how teachers can identify, design and implement *new* approaches for student learning to address weaknesses in student knowledge and understanding.

The findings of this study are consistent with previous research on benchmark testing and extend the extant knowledge of essential factors that promote teachers' formative use of test results. These include the alignment of benchmark tests with the content of curriculum pacing guides as well as the content, format and tools associated with the state exam; test and item quality; and the critical need to create opportunities for teachers to review and discuss the results with colleagues. However, the study findings are limited by the nature of the design and data collection methods. While it was our impression that teachers were very candid in the focus group sessions, the extent to which the group dynamic influenced the accuracy or bias of the information we obtained is unclear. The study is also limited by the self-reported nature of the data collection; focus group sessions were not supplemented by other more direct measures of teachers' use of benchmark test results such as classroom observations or document reviews which may have provided a more comprehensive picture of how benchmark test results are informing the instructional process. Within this context, our findings suggest that future research in this area incorporate factors identified here as being important to successful implementation and use of benchmark test results to improve student learning.

Figure 1. Factors Influencing the Use of Benchmark Test Results


Professional Development
Test and Item Development
How is item quality kept strong and alignment verified?
Test Administration
Online? How much time is taken from instruction?
Scoring and Reporting
Automated? Flexible? Teacher-controlled?
Analysis of Results
Group or individual? Formal or informal? Students motivated? Item transparency?
Use of Results
Re-teaching strategies. For individual students or groups of students? For teacher evaluations? Focus on borderline students?
Expectations
Source
• District
• Principal
• Teachers
Nature of
• Strength
• Consistency

Conclusion

It is clear that school districts across the nation and those at the focus of this study are putting significant resources into benchmark testing, and there is a need to identify factors that enhance the usefulness of the test results. While teachers' perceptions about benchmark testing are mixed and under many of the school contextual conditions described, the benefits of benchmark testing programs may not be worth the financial and time investment required for implementation. However, some results suggest that under the right conditions, benchmark testing may serve a meaningful formative purpose. These conditions include time for teachers to meet, discuss, and collaborate based on test score information; district and school administrations that have adopted supportive rather than punitive paradigms; implementing quality benchmark assessments that are well-aligned to curriculum pacing guides and the state exam as well as produce information that is perceived as value-added. School districts need to consider new and innovative ways to support teachers in using benchmark testing data in ways that are formative. Several directions that could lead to a more powerful formative impact of benchmark test results on instruction and student learning include developing creative schedules, that incorporate time during the school day for teachers to collaborate and provide small group or individualized remediation; evaluating the need for multiple and overlapping assessment systems that increasingly draw time away from instruction; and directing resources to develop high-quality benchmark assessments that are regarded as yielding credible information.

References

Brookhart, S. M. (2007). Expanding views about formative classroom assessment: A review of the literature. In J. H. McMillan (Ed.). *Formative classroom assessment: Theory into practice.* New York: Teachers College Press.

Brown, R. S., & Coughlin, E. (2007). *The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region.* Washington, DC: U.S. Department of Education, Institute of Educational Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Available http://ies.ed.gov/ncee/edlabs.

Brunner, C., Fasca, C., Heinze, J., Honey, M., Light, D., & Mandinach, E. (2005). Linking data and learning: The grow network study. *Journal of Education for Students Placed at Risk, 10*(30, 214-267.

Christman, J. B., Neild, R. C., Bulkley, K., Blanc, S., Lui, R., Mitchell, C., & Travers, E. (2009). *Making the most of interim assessment data: Lessons from Philadelphia.* Philadelphia: Research for Action.

Goertz, M., Olah, L., & Riggan, M. (2009). *Can interim assessments be used for instructional change?* CPRE Policy Briefs: Reporting on Issues and Research in Education and Finance. Available http://www.cpre.org/images/stories/cpre_pdfs/rb_51_role%20policy%20brief_final%20web.pdf.

Henderson, S., Petrosino, A. ,Guckenburg, S. & Hamilton, S.(2008). *A second follow-up year for "Measuring How Benchmark Assessments Affect Student Achievement".* (REL Technical Brief, REL 2008-No. 002). Regional Educational Laboratory Northeast & Islands.

Kerr, K., Marsh, J., Ikemoto, G., Darilek, H. & Barney, J. (2006). Strategies to promote data use of instructional improvement: Actions, outcomes and lessons from three urban districts. *American Journal of Education, 112*, 496-520.

Krueger, R. & Casey, M. (2009). *Focus groups: A practical guide for applied research* (4th ed.). Thousand Oaks, CA: Sage Publications.

Lachat, M. & Smith, S. (2005). Practices that support data use in urban high schools. *Journal of Education for Students Placed at Risk, 10*(3), 333-349.

Marsh, J., Pane, J. & Hamilton, L. (2006). *Making sense of data-driven decision making in education: Evidence from recent RAND research.* Washington, DC: RAND Corporation.

McMillan, J. H. (2007). Formative classroom assessment: The key to improving student achievement. In J. H. McMillan (Ed.). *Formative classroom assessment: Theory into practice.* New York: Teachers College Press.

McMillan, J.H. (2010). The practical implications of educational aims and contexts for formative assessment. In H.L. Andrade & G.J. Cizek (/Eds.), *Handbook of formative assessment.* New York: Routledge (pp. 41-58).

Murnane, R., Sharkey, N., & Boudett, K. (2005). Using student-assessment results to improve instruction: Lessons from a workshop. *Journal of Education for Students Placed at Risk, 10*(3), 269-280.

Nelson, M. & Eddy, R. (2008). Evaluative thinking and action in the classroom. In T. Berry & R. Eddy (Eds.). *Consequence of No Child Left Behind for Educational Evaluation: New Directions for Evaluation, 177*, 37-46.

Niemi, D., Wang, J., Wang, H, Vallone, J., & Griffin, N. (2007). *Recommendations for building a valid benchmark assessment system: Second report to the Jackson Public Schools.* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational measurement: Issues and practice*, *28*(3), 5-13.

Peterson, J. (2007). Learning facts: The brave new world of data-informed instruction. *Education Next,* *7*(1), 36-42.

Popham, W. J. (2008). *Transformative assessment.* Alexandria VA: Association for Supervision and Curriculum Development.

Supovitz, J. & Klein, V. (2003). *Mapping a course for improved student learning: How innovative schools systematically use student performance data to guide improvement.* Philadelphia, PA: Consortium for Policy Research in Education: University of Pennsylvania Graduate School of Education.

Symonds, K. (2004). *After the test: Closing the achievement gaps with data.* San Francisco, CA: Learning Point Association and Bay Area School Reform Collaborative.

Trimble, S., Gay, A. and Matthews, J. (2005). Using test score data to focus instruction. *Middle School Journal, 36*(4), 26-32.

Vogel, L., Rau, W., Baker, P., Ashby, D. (2006). Bringing assessment literacy to the local school: A decade of reform initiatives in Illinois. *Journal of Education for Students Placed at Risk, 11*(1), 39-55.

Wayman, J. & Cho, V. (2009). Preparing educators to effectively use student data systems. In T. Kowalksi & T. Lasley, II. (Eds.), *Handbook of data-based decision making in education* (pp.89-104). New York, NY: Routledge.

Wiliam, D., & Leahy, S. (2007). A theoretical foundation for formative assessment. In J. H. McMillan (Ed.). *Formative classroom assessment: Theory into practice*. New York: Teachers College Press.

Wohlstetter, P., Datnow, A. & Park, V. (2008). Creating a system for data-driven decision-making: Applying the principal-agent framework. *School Effectiveness and School Improvement, 19*(3), 239-259.

Yeh, S. (2006). High-stakes testing: Can rapid assessment reduce the pressure? *Teachers College Record, 108*(4), 621-661.

Appendix A

Teacher Focus Group Protocol

I. INTRODUCTION[1]:

Good afternoon and welcome to our focus group session scheduled for ______ (date). Thank you for taking the time to participate today. My name is __________ and I am from the School of Education at VCU. Where we are working on a project designed to explore teachers' views on benchmarking testing and how the results are or can be used for formative assessment. Our session today will last about 1 hour.

II. FOCUS GROUP SESSION:

The purpose of the focus group session is to find out about your perceptions and use of benchmark testing information. Today we will be discussing benchmark testing generally and then your thoughts on the ways you use the results of benchmark tests for shaping your instruction, assessment, and work with students.

Before we begin, we will review some guidelines that will help the session run smoothly. We will be recording the session so that we can accurately capture all of your comments; it is helpful if you speak one at a time. Also, we want to assure you of complete confidentiality, so please use your first name only during today's session. In the written summaries of the session no names will be attached to comments. It is also important that you assure each other of complete confidentiality by not sharing any of the information discussed in this session.

I am interested in all of your viewpoints – both positive and negative. When responding to the questions, please omit specific names of individuals, such as other colleagues or students in your classes. Each time you begin your response to the focus group questions please start by stating your first name.

TOPIC 1: NATURE OF BENCHMARK TESTS AND TEST RESULTS

1. What is your division's benchmark testing policy and process?
   - How long have the tests been used?
   - Who developed(s) the test?
   - How is it administered?

[1] Adapted from Kruger, R. (1994). *Focus groups: A practical guide for applied research* (2nd edition). Sage Publications: Thousand Oaks, CA.

2. What type of data or information do you receive after students complete the test?

    - When do you receive the data?

    - How are the data presented (e.g., class averages, individual student, tables, individual item, means, reporting category)?

3. Does any other information accompany the test results?

4. Does the data help to identify students' weaknesses/strengths?

TOPIC 2: EXPECTATIONS FOR USING BENCHMARK TEST INFORMATION

1. What are division policies or expectations concerning how teachers should use the test results/information?

2. Are these expectations different than those in your school or department/team?

TOPIC 3: INSTRUCTIONAL USES OF BENCHMARK TEST INFORMATION

1. What do you typically do when you receive the results?

2. To what extent do the results influence your instruction? If so, what types of changes typically occur?
   Prompt: Describe an example/scenario of how you used the test results to inform or modify your instruction.

   Prompt: What specific changes to instruction were made? Did you modify the methods or strategies – more teaching, re-teaching, review the alignment between the cognitive process or verb expressed in the standard and the instructional strategy?

3. Do the test results ever tell you something you didn't already know about student performance or proficiency?

4. Are there any barriers to using the information? Supports?
   Prompt: For example, do you utilize your school or division instructional specialist?

   Prompt: What type of training would be helpful? (e.g., interpreting test scores or how to make use of the test score information)

Prompt: What type of materials would be helpful? (e.g., manipulatives, instructional books, websites)

TOPIC 4: VIEWS ON BENCHMARK TESTING

1. Do the results match what you have seen in your classroom assessments?

2. How could the benchmark test results be presented in a more useful way?

3. What types of training would you recommend to help teachers use the results more effectively?

4. What improvements or recommendations do you have regarding the practice of benchmark testing?

5. Should use of the benchmark tests continue?

III. CLOSING:

Option 1: *Time Still Remaining*: Before we end the session, are there any other comments that you have or topics that we missed in our discussion? Thank you for your time and participation.

Option 2: *Time is Up*: If, after today's session, you think of any other comments or topics that were missed please feel free to contact me (e.g., focus group facilitator) by email. Thank you for your time and participation.

Appendix B

Data Analysis Codes and Frequencies

Key categories	Subcategories	Codes	Frequency
Benchmark Testing Policy (POL)			
POL: Length of testing program	# of years	POL-LENG	6
POL: Test development process	County developed	POL-DEV-CO	33
	Externally developed	POL-DEV-EXT	2
	Use of released SOL items	POL-DEV-SOL	6
	Use of pilot test	POL-DEV-PLT	4
	Teacher constructed (use of database or item pool)	POL-DEV-DB	9
POL: Alignment	SOL content	POL-ALIGN-CONT	17
	SOL format	POL-ALIGN-FORM	17
	High difficulty level	POL-ALIGN-HIDIFF	3
	Low difficulty level	POL-ALIGN-LODIFF	6
POL: Administration process	Paper-pencil format	POL-ADMIN-PAP	7
	Online administration	POL-ADMIN-ONL	16
	Quarterly administration	POL-ADMIN-QUART	10
	Untimed	POL-ADMIN-UNTIME	4
POL: Grades	Results can be used as a grade	POL-GRD-INC	19
	Results cannot be used as a grade	POL-GRD-NOINC	9
Receipt of Test Results (RES)			
RES: Timing	Immediate	RES-TIM-IMM	12
	Couple of weeks	RES-TIM-WEEKS	1
RES: Level of reporting/analysis	District level	RES-LEV-DIS	7
	School level	RES-LEV-SCH	4
	Grade/Team/Department level	RES-LEV-GRD/TEAM/DPT	4
	Class level	RES-LEV-CLASS	20
	Student level	RES-LEV-STUD	22
RES: Type of data (Refers to type of results).	Measures of central tendency	RES-TYPE-CEN	1
	Disaggregated by strand	RES-TYPE-STR	8
	By SOL standard	RES-TYPE-SOL	11
	AYP subgroups	RES-TYPE-AYP	1
	By item	RES-TYPE-ITEM	16
RES: Analysis of data (Refers to use of results).	Measures of central tendency	RES-ANL-CEN	0
	Disaggregated by strand	RES-ANL-STR	11
	By SOL standard	RES-ANL-SOL	2
	AYP subgroups	RES-ANL-AYP	2
	By item	RES-ANL-ITEM	22
	AYP performance categories	RES-ANL-PERF	0

	Group discussion	RES-ANL-GROUP	17
RES: Format	Graphs Raw scores Percentiles	RES-FORM-GRAPH RES-FORM-RAW RES-FORM-PERC	1 1 6
RES: Who provides	District Principal Independent access to database	RES-WHO-DIS RES-WHO-SCHOOL RES-WHO-IND	1 9 22
Expectations for Teacher Use (EXP)			
EXP: District expectations	Overall district expectations	EXP-DIS	6
EXP: Principal or building expectations	Identification of weak students Identification of curricular areas to review Communicated to students Review test with students Adjust instruction Re-teach content Analyze discuss the results	EXP-PRIN-IDSTUD EXP-PRIN-IDREV EXP-PRIN-COMSTUD EXP-PRIN-REVTEST EXP-PRIN-INS EXP-PRIN-RETCH EXP-PRIN-ANL	6 9 1 5 5 5 8
EXP: Departmental or team use	Allocation of resources between classes	EXP-TEAM-RES	0
Instructional Uses of Results (INS)			
INS: Identify students' strengths/weakness	Student self-assessment Identify student strengths Identify student weaknesses Differentiate instruction	INS-STUD- STUDASSESS INS-STUD-STR INS-STUD-WEAK INS-STUD-DIFF	8 16 49 5
INS: Identify curricular areas/topics	Identifies curricular areas in need of review or to re-teach Test preparation Informs subsequent instruction: upcoming units next academic year	INS-CURR-REV INS-CURR-PREP INS-CURR-SUB-UNIT INS-CURR-SUB-YEAR	46 18 6 4
INS: Determine approach to remediation or re-teaching	Individual Small group Entire class Identify assignment to school programs: during the day after school weekend	INS-REM-IND INS-REM-GRP INS-REM-CLA INS-REM-PRO-DAY INS-REM-PRO-AFT INS-REM-PRO-WKEND	20 13 18 6 4 0
Supports for using test results (SUP)			
SUP: Resources	Instructional resources Time during the day Computer database Quick turnaround on results	SUP-RES-INS SUP-RES-TIME SUP-RES-DB SUP-RES-QCK	2 1 10 4

	Professional development	SUP-RES-PD	5
SUP: Specialists	District specialists	SUP-SPEC-DIS	0
	School specialists	SUP-SPEC-SCH	8
SUP: Principal	Motivation	SUP-PRIN-MOT	7
	Resources	SUP-PRIN-RES	3
	Identify students for remediation	SUP-PRIN-IDSTUD	4
SUP: Environment	School	SUP-ENV-SCH	10
	Department/team	SUP-ENV-DEPT/TEAM	2
SUP: Accountability	Parent accountability	SUP-ACC-PAR	4
	Student accountability/motivation	SUP-ACC-STUD	20
	Teacher accountability	SUP-ACC-TEACH	34
Obstacles/barriers to using test results (OBS)			
OBS: Pacing	Pacing	OBS-PAC	7
	Alignment with pacing guide	OBS-PAC-ALIGN	26
	Need to cover content	OBS-PAC-CON	20
OBS: Time	Teachers' time required to generate results	OBS-TIME-TEACH	14
	Instructional time for admin	OBS-TIME-INS	20
OBS: Student	Student	OBS-STUD	3
OBS: Scoring error	Error in scoring	OBS-ERR-SCOR	1
OBS: Quality	Poorly constructed/insufficient items	OBS-QLTY-ITEM	21
	Overall test	OBS-QLTY-OVR	15
	Reading passages	OBS-QLTY-READ	9
OBS: Technology	Technology	OBS-TECH	18
Utility of Benchmark Testing (UTL)			
UTL: Additional information provided	Indicator of student performance	UTL-ADD-IND	19
	Confirmation of professional judgment/other classroom assessment	UTL-ADD-ASS	12
	Provides additional information related to student population	UTL-ADD-STUD	23
	Provides evidence of student improvement	UTL-ADD-IMP	5
	Varies by subject	UTL-ADD-VAR	7
Value of Benchmark Testing (VAL)			
VAL: Overall	Positive	VAL-OVER-POS	56
	Negative	VAL-OVER-NEG	103
VAL: Impact	Impact on student achievement	VAL-IMP-ACHIEVE	3
	Impact on classroom assessment	VAL-IMP-ASS	5
Recommendations (REC)			

REC: Improve practice of formative use of test results	Improved quality of tests	REC-PRAC-TEST QLTY	5
	Time during the instructional day	REC-PRAC-TIME	4
	Professional development:		
	data base program	REC-PRAC-PD-DB	5
	interpreting test information	REC-PRAC-PD-INT	1
	Identifying actionable information	REC-PRAC-INFO	0
REC: Make results more useful/helpful	Comparison reports across grade level and county	REC-RES-COMP	9
FLAG	Significant comment – no code	FLAG	58

Appendix C

Code Dictionary

Codes	Definitions
Benchmark Testing Policy	
POL-LENG	How long the benchmark testing program has been implemented in the district.
POL-DEV-CO	Benchmark tests are developed by the county by content area specialists, teachers, other personnel.
POL-DEV-EXT	Benchmark tests are developed by outside company.
POL-DEV-SOL	Benchmark tests consist in part of released SOL items.
POL-DEV-PLT	Benchmark tests were piloted.
POL-DEV-DB	Benchmark tests are constructed by teachers using a database
POL-ALIGN-CONT	Benchmark test items are aligned with SOL content.
POL-ALIGN-FORM	Benchmark test items are formatted like SOL test items.
POL-ALIGN-HIDIFF	Benchmark tests are more difficult than SOL tests.
POL-ALIGN-LODIFF	Benchmark tests are less difficult than SOL tests.
POL-ADMIN-PAP	Administered in paper/pencil format.
POL-ADMIN-ONL	Administered online.
POL-ADMIN-QUART	Administered on a quarterly (every nine weeks) basis.
POL-ADMIN-UNTIME	Benchmark tests are untimed like SOL tests.
POL-GRD-INC	School division/school has a policy allowing teachers to include benchmark test results in student grades.
POL-GRD-NOINC	School division/school has a policy prohibiting teachers to include benchmark test results in student grades.
Receipt of Test Results	
RES-TIM-IMM	Teachers have access to the results shortly (within 1 week after test administration) after the students complete the test.
RES-TIM-WEEKS	There is some delay (greater than 2 weeks after test administration) between administration and teacher's receipt of results.
RES-LEV-DIS	Results are reported/analyzed at the district level.
RES-LEV-SCH	Results are reported/analyzed at the school level.
RES-LEV-GRD/TEAM/DPT	Results are reported/analyzed at the grade/team/department level.
RES-LEV-CLASS	Results are reported/analyzed at the class level.
RES-LEV-STUD	Results are reported/analyzed at the student level.
RES-TYPE-CEN	Measures of central tendency are reported.
RES-TYPE-STR	Results disaggregated by SOL strand are reported.
RES-TYPE-SOL	Results disaggregated by SOL standard are reported.
RES-TYPE-AYP	Results disaggregated by AYP subgroup are reported.
RES-TYPE-ITEM	Item level results are reported.

RES-ANL-CEN RES-ANL-STR RES-ANL-SOL RES-ANL-AYP RES-ANL-ITEM RES-ANL-PERF RES-ANL-GROUP	Teachers analyze results using measures of central tendency. Teachers analyze results disaggregated by SOL strand. Teachers analyze results disaggregated by SOL standard. Teachers analyze results disaggregated by AYP subgroup. Teachers analyze results by item statistics. Teachers analyze results disaggregated by AYP performance categories. A combination of teachers, specialists, and administrators analyze results as a group.
RES-FORM-GRAPH RES-FORM-RAW RES-FORM-PERC	Reporting of results include graphs. Reporting of results include raw scores. Reporting of results include percentiles.
RES-WHO-DIS RES-WHO-SCHOOL RES-WHO-IND	Results are provided by district level personnel. Results are provided by school leadership. Teachers have direct access to results through data management system.
Expectations for Teacher Use	
EXP-DIS	District policies establish how schools/teachers should use results.
EXP-PRIN-IDSTUD EXP-PRIN-IDREV EXP-PRIN-COMSTUD EXP-PRIN-REVTEST EXP-PRIN-INS EXP-PRIN-RETCH EXP-PRIN-ANL	Principals expect teachers to use results to identify and remediate weak students. Principals expect teachers to use results to inform their instruction. Principals communicate test expectations to students. Principals expect teachers to review tests with students. Principals expect teachers to use results to adjust instruction. Principals expect teachers to reteach content based on test results. Principals expect teachers to analyze and discuss test results.
EXP-TEAM-RES	Departments/teams use results to allocate resources.
Instructional Uses of Results	
INS-STUD-STUDASSESS INS-STUD-STR IND-STUD-WEAK INS-STUD-DIFF	Results are used by students to identify personal strengths and weaknesses. Results are used by teachers to identify student strengths. Results are used by teachers to identify student weaknesses. Results are used by teachers to differentiate instruction.
INS-CURR-REV INS-CURR-PREP INS-CURR-SUB-UNIT INS-CURR-SUB-YEAR	Results are used to identify curricular areas to review or to inform reteaching. Results are used to inform test preparation strategies for students. Results are used to inform instruction in upcoming units. Results are used to inform instruction in the next academic year.
INS-REM-IND INS-REM-GRP INS-REM-CLA INS-REM-PRO-DAY INS-REM-PRO-AFT	Results are used to target instruction/remediation to individual students. Results are used to target instruction/remediation to small groups. Results are used to target instruction/remediation to entire class. Results are used to assign students to remediation during the day.

INS-REM-PRO- WKEND	Results are used to assign students to remediation after school. Results are used to assign students to remediation over the weekend.
Supports for Using Test Results	
SUP-RES-INS SUP-RES-TIME SUP-RES-DB SUP-RES-QCK SUP-RES-PD	Instructional resources support teachers' use of test results. Time during the day for planning supports teachers' use of test results. Access to/ a data management system to sort and retrieve results supports teachers' use of results. Teachers' timely receipt of results supports their use. Professional development for teachers supports teachers' use of test results.
SUP-SPEC-DIS SUP-SPEC-SCH	District specialists provide support that enables teachers' use of results. School specialists provide support that enables teachers' use of results.
SUP-PRIN-MOT SUP-PRIN-RES SUP-PRIN-IDSTUD	Principal provides teachers and/or students with motivation. Principal provides teachers with resources which enables use of results. Principal assists teachers in identifying students for remediation.
SUP-ENV-SCH SUP-ENV-DPT/TEAM	Elements of the school environment support teachers' use of results. Elements of the department/team environment support teachers' use of results.
SUP-ACC-PAR SUP-ACC-STUD SUP-ACC-TEACH	Parent accountability for student performance on benchmark tests. Student accountability/motivation for student performance on benchmark tests. Teacher accountability for student performance on benchmark tests.
Obstacles/Barriers to Using Test Results	
OBS-PAC OBS-PAC-ALIGN OBS-PAC-CON	General comment about pacing. Disconnect between content of the test and content of instruction or the pacing guide. Teachers have difficulty covering the content of the pacing guide.
OBS-TIME-TEACH OBS-TIME-INS	Time required for teachers to score tests and generate results inhibits their use. Instructional time lost to administration of tests.
OBS-STUD	Student as an obstacle to teachers' utility and value of benchmark test results.
OBS-ERR-SCOR	Scoring errors inhibit teacher use of tests results.
OBS-QLTY-ITEM OBS-QLTY-OVR OBS-QLTY-READ	Poorly constructed or insufficient items inhibit use of test results. Overall quality of the test inhibits use of test results. Content and format of reading passages inhibit use of test results.
OBS-TECH	Technology as an obstacle in test administration and use of test results.
Utility of Benchmark Testing	

UTL-ADD-IND	Benchmark testing provides an indicator (predictor) of student performance.
UTL-ADD-ASS	Benchmark testing provides teachers with confirmation of professional judgment or other classroom assessments.
UTL-ADD-SOL	Benchmark testing provides teachers a way to prepare students for the SOL tests.
UTL-ADD-STUD	Benchmark testing provides teachers with additional information about their students.
UTL-ADD-IMP	Benchmark testing provides teachers with evidence of student improvement.
UTL-ADD-VAR	Utility of benchmark testing varies by subject area.
Value of Benchmark Testing	
VAL-OVER-POS	Teacher describes an overall positive value of benchmark testing.
VAL-OVER-NEG	Teacher describes an overall negative value of benchmark testing.
VAL-IMP-ACHIEVE	Benchmark testing impacts student achievement.
VAL-IMP-ASS	Benchmark testing impacts classroom instruction.
Recommendations	
REC-PRAC- TEST QLTY	Improving the quality of the tests (scoring errors, administration errors, test items) may improve practice of formative use of test results.
REC-PRAC-TIME	Additional time during the instructional day may improve practice of formative use of test results.
REC-PRAC-PD-DB	Professional development on the data management system may improve practice of formative use of test results.
REC-PRAC-PD-INT	Professional development on how to interpret test information may improve practice of formative use of test results.
REC-PRAC-INFO	Aid in identifying actionable information may improve practice of formative use of test results.
REC-RES-COMP	Reports comparing results across grade level and county may make results more useful/helpful.
FLAG	Significant comment – no code