CRESST REPORT 807

HOW MIDDLE SCHOOL MATHEMATICS TEACHERS USE INTERIM AND BENCHMARK ASSESSMENT DATA

OCTOBER, 2011

Lorrie A. Shepard
Kristen L. Davidson
Richard Bowman

National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

How Middle-School Mathematics Teachers Use Interim and Benchmark Assessment Data

CRESST Report 807

Lorrie A. Shepard and Kristen L. Davidson
CRESST/ University of Colorado at Boulder

Richard Bowman
CRESST/ Pardee RAND Graduate School

October, 2011

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences (IES), U.S. Department of Education.

To cite from this report, please use the following as your APA reference: Shepard, L., Davidson, K., & Bowman, R. (2011). *How middle-school mathematics teachers use interim and benchmark assessment data.* (CRESST Report 807). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

HOW MIDDLE-SCHOOL MATHEMATICS TEACHERS USE INTERIM AND BENCHMARK ASSESSMENT DATA

Lorrie A. Shepard and Kristen L. Davidson
CRESST/University of Colorado Boulder

Richard Bowman
CRESST/Pardee RAND Graduate School

Abstract

In 2001, the enactment of the No Child Left Behind Act intensified the pressure on school districts to raise test scores, close achievement gaps, and turn around low-performing schools. In response, a large number of school districts have adopted interim or benchmark assessments to be administered periodically throughout the school year in anticipation of annual state tests. This report focuses on middle-school mathematics teachers' uses of interim and benchmark assessment results. We present findings from two-stage interviews with 30 teachers in seven districts across two states. While teachers' uses of assessment information varied, few gained substantive insights about students' mathematical understanding. Instead, teachers most frequently retaught standards or items with the lowest scores and focused on procedural competence. Although many teachers expressed an interest in using assessment results to inform instruction, they reported minimal professional development to this end, and often had a different understanding regarding the intended use of the assessments than did district leaders.

Introduction

Background

Since their inception, standards-based reforms and test-based accountability have relied on the idea that test data should be used to improve instruction and increase student achievement. In addition to creating incentives for change through accountability pressure, achievement test scores are expected to provide essential information about what is working and what is not so that educators can act to make needed improvements. Originally this theory of action (with or without intervening variables, such as teacher professional development and improved instructional practices) focused on the use of end-of-year, summative tests (Elmore & Rothman, 1999).

In 2001, however, the enactment of No Child Left Behind (NCLB) dramatically intensified the pressure on school districts to raise test scores, close achievement gaps, and turn around low-performing schools. In response, a large number of school districts have adopted interim or benchmark assessments to be administered periodically throughout the school year in

anticipation of annual state tests. In a recent survey of large urban school districts, 82% reported that they had instituted some form of interim assessment and 69% of these had done so following the passage of NCLB (Burch, 2010).

Perie, Marion, and Gong (2009) offered a framework for considering how interim assessments might be used as part of a comprehensive assessment system and defined interim assessments based, in part, on their middle-range time-scale location somewhere between once-per-year state accountability tests and minute-by-minute formative assessments embedded within classroom instructional activities. Typically, interim assessments are similar in format to accountability tests and results can be aggregated by classroom and school. They can also serve formative instructional purposes, but only if they are substantively aligned with local curricula and are timed to allow teachers to adapt instruction.

Given that the prevalent use of interim assessments appeared to spring up overnight, they lack a research base of their own. Although some instruments, such as the Northwest Evaluation Association's (NWEA) *Measures of Academic Progress* (MAP®), have been around for decades, few studies have been conducted to examine the technical adequacy of interim assessments or to evaluate their effects on teaching and student learning. Proponents and product advertisers have relied instead on two distinct research literatures, corresponding to the two ends of the time-scale continuum identified by Perie et al. (2009): research on instructionally grounded formative assessment and research on school- and district-level data-based decision making. Both literatures are summarized here briefly because either could reflect the intentions of district officials in adopting interim assessments or the understandings of individual classroom teachers who are ultimately responsible for using the information gained from interim assessments.

Purpose of the Study

This report focuses on middle-school mathematics teachers' uses of interim and benchmark data. It relies primarily on teacher interviews conducted in a two-stage interview process as described in the Methods section. The work reported here was part of a larger study that addressed both interim assessments and classroom formative assessment practices. The full study design included interviews with district-level administrators—superintendents or deputy superintendents, mathematics coordinators, assessment directors, and professional development directors—as well as school-level principals. Findings from analyses of district-level data are reported in Davidson and Frohbieter (2011) but are excerpted here in comparison to teachers' understandings of district expectations for the use of their respective testing programs. As part of the sampling strategy described in the Methods section, some districts were selected to represent instances of district-wide implementation of formative assessment and professional development,

which could be thought of as an alternative to investing in an interim assessment system. Those cases are described in Frohbieter, Greenwald, Stecher, and Schwartz (2011).

Our research questions specific to teachers' use of interim assessments were as follows:

1. What is the context for teachers' use of interim assessments? For example: a) What do teachers believe are teachers' understandings of their school district's intentions for the use of interim assessments? b) What types of professional development do teachers receive to help them use the assessments?
2. What information do teachers gain about students' understanding of mathematics from various forms of interim assessments?
3. How do teachers use the information gained from interim assessment results?

Although not explicit as a separate research question, a clear goal of the study was to try to get beyond the theories, advertisements, and rhetorical claims and identify how various assessments were actually being used. This was more than merely a distinction between ideal and actual practice; rather, it was aimed at a different level of specificity. We wanted to gather specific examples of each type of use instead of rely on vague generalizations. When a teacher said, for example, "I use the data to adapt instruction"—could we glean specific examples of the particular information that the teacher learned from the assessment? How did he or she decide that instruction needed to be modified? What adaptation was actually made?

Review of the Literature

Research on Formative Assessment

Research on formative assessment was synthesized most famously by Black and Wiliam (1998a), who brought together disparate bodies of work addressing feedback, motivation, self-assessment, classroom discourse, the nature of teacher questioning, and so forth. By citing typical effects or formal meta-analyses[1] from these various research literatures, Black and Wiliam (1998b) concluded that formative assessment had the potential to increase student learning by .4 to .7 standard deviations; these were far greater gains than typical educational interventions. If widely implemented, they noted, for example that an effect size of .7 standard deviations would be sufficient to raise a country in the middle of the pack in international comparisons (such as the U.S.) to among the top five nations in the world. However, what is also clear from the many studies cited is that effective formative assessment processes cannot be

[1] It is important to note that Black and Wiliam (1998a) did not conduct a meta-analysis across the quite varied literatures considered in their review (see a critique by Bennett (2009)). Rather their famously quoted effect sizes came from a few exemplary studies or from meta-analyses of distinct subareas such as studies of feedback (Kluger & DeNisi, 1996).

implemented merely by adopting a single strategy or instrument. For instance, in a meta-analysis of 131 controlled studies examining the effects of feedback, Kluger and DeNisi (1996) found that the types of feedback given in one-third of the studies produced negative outcomes.

Making sense of how formative assessment works to improve learning, when it works, requires a complex weaving together of both learning and motivational theories. For example, feedback is more effective when it is focused on features of the task (rather than seeming to evaluate the person as good or bad) and when it provides the learner with information about how to improve (see, for example, Bangert-Drowns, Kulik, Kulik, & Morgan, 1991). Indeed, in Sadler's (1989) early model of formative assessment, the fundamental purpose of feedback is not served by a letter grade or right-wrong scores; rather, it must engage the student in such a way that the student can come to a shared understanding of the teacher's concept of quality and can ultimately self-monitor progress toward that shared goal. Shepard et al. (2005) pointed out the close parallels between this model of formative assessment processes and instructional scaffolding, whereby learners are supported while developing increasingly internalized and independent demonstrations of mastery. As many have noted, this research-based conception of formative assessment is deeply embedded in instructional interactions; therefore, formative assessment cannot be neatly separated from good instructional practices aimed at developing students' deep knowledge and conceptual understanding.

When interim assessments were first widely promulgated as tools for addressing the demands of NCLB, providers called them "formative assessments;" the Black and Wiliam (1998a, 1998b) review was cited as evidence of their efficacy. This use of the term formative assessment was controversial (Chappuis, 2005), because quarterly administrations of a formal test could not enable real-time adaptation of instruction as implied by research studies. Wiliam (2004) instead called them "early-warning summative tests" (p. 4). Shepard (2008) drew the distinction between formative *assessment* with its focus on moving learning forward during instruction and formative *program evaluation* for which both state tests and district interim tests could be used. A state collaborative sponsored by the Council of Chief State School Officers (CCSSO) attempted to resolve the confusion by issuing a definition (McManus, 2008):

> 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 (p. 3).

Similarly, an early version of the Perie et al. (2009) article and stories in *Education Week* helped to clarify the distinctions and established *interim* and *benchmark* as more appropriate labels for longer-term, periodic tests.

Research on Data-Based Decision Making

The intention of using standardized test data to improve instructional programs has a long history. Both E.F. Lindquist (1951) and Ralph Tyler (1951) conceived of district testing programs that would be closely tied to ongoing evaluations of curriculum and instruction. Effective schools research led by Ron Edmonds (1979) identified frequent monitoring of student progress as one of the distinguishing characteristics of schools that performed above their statistical predictions. In the early 1990s, the theory of action underlying standards-based reform assumed that schools would monitor student progress against the standards and make instructional changes to address any shortfall (Elmore & Rothman, 1999). Currently, various phrases are used in the educational leadership, policy, and measurement literatures to describe this idea of inquiry-driven school improvement, which has intensified following NCLB. This vision of continuous improvement and organizational learning closely parallels and draws from business and total quality management (TQM) practices advanced by Deming (1986). Trained as a statistician, Deming's principles for achieving TQM rested importantly on the gathering of statistical evidence of quality as well as a philosophical commitment to it.

Researchers have sometimes noted that data use is associated with the development of a more professional and collaborative culture (Feldman & Tung, 2001) – what Boudett, City, and Murnane (2005, p. 95) called "a culture of improvement [based] [on] [a] habit of inquiry." In some cases, professional development efforts focused on data interpretation and data use are explicitly designed to foster professional learning communities. While each individual teacher could use accountability test data to analyze and revise his or her own teaching program, most of the theoretical frameworks associated with data-driven instructional improvement focus on schools as the locus of collaborative effort and principals as key leaders. For example, Halverson, Grigg, Prichett, and Thomas (2007) describe data-driven instructional systems in which six elements—data acquisition, data reflection, program alignment, instructional design, formative feedback, and test preparation—work together to "link the results of summative testing to formative information" (p. 163). To avoid confusion, we should clarify that Halverson et al. (2007), use the term "formative feedback" to refer to evidence at the school level in regards to how well program initiatives are working; it is their data reflection step that more likely involves the use of formative feedback to further individual student learning.

Research summaries on data-based decision making also cite several persistent barriers to effective implementation. Until recently, access to data was the most obvious obstacle, and even with myriad new software products, adequate technology, professional development, and effective leadership, it may still be "the exception rather than the rule" (Wayman, 2005, p. 296). Researchers studying standards-based reforms have persistently found that educators often have

difficulty drawing appropriate links between assessment outcomes and instructional practices (Elmore & Rothman, 1999; Herman & Gribbons, 2001). In an NSF-funded study of data-driven decision making, Mandinach, Honey, Light, and Brunner (2008) found that school administrators tended to be more adept at identifying broad patterns of strengths and weaknesses from high-stakes accountability data; whereas, teachers preferred to use multiple sources of data, including teacher-created assignments, and to focus on individual students rather than looking for classroom-wide patterns.

Methods

Sampling of Assessments, Districts, and Teachers

Based on the distinctions among assessments made by Shepard (2005) and Perie et al. (2009), we used a proximity continuum for the full study to characterize the closeness or distance of various assessment types to instruction. As shown in Figure 1, interim and benchmark assessments on the left-hand side are administered periodically and may be more or less coordinated with a particular curriculum sequence. We conducted web surveys to identify the commercial products available, contacted district assessment directors by phone and through a list-serve, and noted interim assessments mentioned frequently in *Education Week* articles.

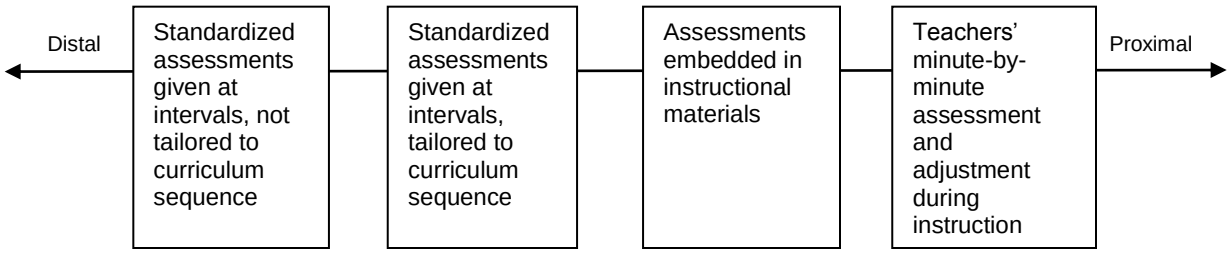

Figure 1. A continuum of interim and formative assessments.

Our sampling process was purposive and also one of convenience (Maxwell, 2005). We sought to include assessments that represented each type along the proximity continuum and that were in prevalent use nationally and in the three geographic regions in which research team members were located. Participation was first sought at the district level. The district administrator responsible for middle school mathematics was then asked to identify two schools where the district assessment was being effectively implemented by the principal and teachers. Our goal was not to sample the full range of practices across the district but rather to identify classrooms where it was reasonably likely that the assessments were being used by teachers. The district administrator recruited the principals who in turn were asked to recruit three mathematics teachers to be interviewed. Teachers received $50 gift certificates for a bookstore in recognition

of their time spent preparing for and participating in the interviews. In some cases we were not successful in recruiting the number of schools and teachers sought for each district. The final sample for the districts with interim or benchmark assessments is shown in Table 1.

Table 1

Interim and Benchmark Assessments and Number of Teachers from Participating Districts

School district	Number of teachers	Assessments
Adlington	3	District-developed benchmark assessment[a]
Burlington	4	District-developed assessment using items and score reporting by Evans-Newton
Madison	5	District-developed interim assessment using some items from Connected Math
Pittsfield	4	District-developed quarterly assessment using items from Holt textbook
Sinclair	3	District-developed benchmark assessment[a]
Taylor	5	Northwest Evaluation Association Measure of Academic Progress System (NWEA MAP®)
Washington	6	Northwest Evaluation Association Measure of Academic Progress System (NWEA MAP®)

[a] Two districts, Adlington and Sinclair, also implemented the POWERSOURCE© formative assessment strategy in addition to their district-developed benchmark assessments.

Teacher Interview Protocols

As part of the recruitment process, principals and teachers received a form that explained the purpose of the study, research methods, and what was needed from participating teachers. Once they agreed to participate, teachers were provided with a more detailed consent form. We proceeded with a two-stage interview process in order to ask general questions in the first stage that could then be pursued in greater depth in the second stage. In particular, teachers were asked to bring examples to the second interview of particular uses of assessment data that had been discussed in the first interview.

Interview 1 was conducted in a scheduled phone call of approximately 40 to 60 minutes. Teachers were informed that later in the interview they would be asked about formative assessments that they used in their classroom, both formal and informal, and about assessments they used for grading; yet, to begin, our first set of questions focused on the district's interim or

benchmark assessment. We asked what the district's expectations were in deciding to use the specifically adopted interim assessment and what kind of technical support or professional development had been provided to help teachers learn how to use the system.

We then prompted teachers to elaborate on the following:

- "Please describe, in as much detail as possible, how the (particular interim or benchmark) assessment works in your classroom."
- "What kind of information do you get from it?" "Can you give specific examples of the information you get?" "What do you do with this information?"
- "What has the (particular interim or benchmark) helped you learn about your students?" "How often do these (insights or new information) occur?" "How often are you able to use these (insights) in your teaching?"

Note that the research team was particularly interested in the possibility of substantive insights gained from instructionally based assessments; in fact, "substantive insight" was later used as a coding category. However, we did not want to supply this idea or term to interview participants. Therefore, we attempted to adopt the terms that each individual teacher used in responding to prior questions.

At the end of the first interview, plans were made for the second hour-long interview, which was to be conducted in person approximately two weeks later. We explained that "one goal of our study is to get more specific and concrete about the nature of the help that is provided by different types of formative and interim assessments." We planned to come back to some of the same questions and asked that teachers collect specific examples that could be shared with us. For instance, they might bring printouts from the most recent administration of the district's interim assessment or samples of student work on specific assessment tasks. We asked that they collect examples for a student who needs extra support, a typical student, and a top performing student. Instructions were provided on how to conceal student names and make copies of data printouts and sample assessments with reimbursement if copying charges were an issue. In the second interview, questions like those from the first interview were repeated. Additional questions probed into how this information might vary for low-, middle-, and high-achieving students and whether what was learned from the assessment pertained to individual students or the class as a whole.

Both interviews for each teacher were audio-recorded and transcribed. Assessment artifacts were digitized and linked to the place in the transcript where they were referenced.

Data Analysis

Qualitative coding of the interview transcripts was carried out in several phases using NVivo software. In the first phase, a list of codes and definitions was developed through an extended iterative process. The team of eight researchers first read through teacher cases in an unstructured way looking for themes and patterns of response. Potential codes were identified and supported with illustrative quotations. These codes were then applied to new cases that were read by pairs of researchers and reconciled through discussion of discrepancies before they were shared with the entire team. Using this process, we continually refined our codebook, and created specific inclusion and exclusion criteria for each code in order to achieve consistency among coders. Although new codes could be added at any time, the tendency was for unreliably coded descriptors to be folded into larger categories with the understanding that more specific subcategories would be systematically analyzed in the third phase of analysis. Once the final list of codes was agreed upon, they were applied to all of the teacher cases, again using the process of pair-wise reading and reconciling.

The phase 2 coding list comprised several major categories: *assessment type* (whether the assessment was mandated by the district or was one of several formal or informal classroom level assessment sources); *assessment features* (timing, format, to which students it was given, and other features); *assessment information* (whether the information gained was at the school level, for individual students, the class as a whole, or about the teacher's curriculum and teaching decisions); *assessment use* (whether action was taken with the entire class or with individual students, or was used for grouping, grading, placement, self-assessment, student feedback, or parent-teacher communication); and *assessment philosophy or concerns* (including student reactions, teacher concerns about assessment quality or time constraints, and teachers' beliefs about assessment or instruction). *State test* included any statements about the state test, but most frequently referred to alignment or preparation. *Professional development* included both district-sponsored efforts and teacher-initiated collaboration associated with use of assessment data. A few codes were specific to the interview protocol, such as *district intent* which was used to track responses to the first interview question and other mentions of the district's purpose in adopting the assessment program. All of these codes could overlap within any given transcript segment.

For the final phase of analysis, the master NVivo file was used to generate reports (i.e., exhaustive transcript listings) based on codes and groups of codes using Boolean logic. Interim and formative assessment types were assigned to teams of three to four researchers, but the entire research team convened to agree upon an analogous coding scheme for the assessment information and assessment use codes. Sub-categories were created within each of the second-phase codes to more accurately characterize teachers' practices (Miles & Huberman, 2002).

Findings reported in this paper reflect these more interpretive codes, including: (1) information obtained with regard to *mastery of content standards*, *insight*, and *evaluation of curriculum/evaluation of teaching*; (2) use of information by means of *reteach/review*, *student intervention*, *feedback to students*, and *placement*; and (3) *overall usefulness of assessment information*. Interview transcripts from each teacher were reread; summary tables were created for each teacher using the most representative quotation in each category. Next we developed "read-across" summaries for each category (Miles & Huberman, 2002). We sought to typify the assessment information and uses associated with each code, but we also represented the range of responses captured by that code.

The interim assessment team noted that much of the data were disappointingly vague, and sought to address this concern by distinguishing among superficial users and more "fulsome" users of assessment data. To accomplish this, we noted which types of information and uses were most prevalently mentioned or most emphasized by each teacher, followed by supplemental types of information and uses. In this way, we could characterize the teachers' typical classroom practices with regard to the assessment system. We describe the teachers with more elaborated practices in greater detail, while at the same time recognizing that they are exceptional. We also organized teacher summaries by school and district so that it was possible to see when particular assessment practices were shared or were idiosyncratic to individual teachers.

Findings

District Intent

We asked teachers in six districts what their district's expectations were in deciding to use the interim or benchmark assessment.[2] Teachers' answers were quite general and most prevalently indicated two purposes, which often overlapped in their responses: a) mastery or progress toward standards and b) preparation for the state test. The following quotation is illustrative of the category we called *mastery of standards*:

> The way I understand it, the benchmark assessment was to give us an idea of what the kids have learned so far. And it was also to help us inform our instruction. You know, what did they get, what did they not get, what did we need to cover, where were the strengths, where were the weaknesses, and all of that data was collected, and analyzed, and then sent back to

[2] The question about district intent was omitted for the three teachers in the seventh district where the assessment was being piloted in only some schools; a technical difficulty made the response to this question unavailable for one additional teacher. Therefore, perceptions of district intent are reported for only 26 of the 30 teachers in the study.

> us, and then we could then analyze it even further, based on not only the particular kids, but the overall class ***(Sinclair SD, Rose).***

Teachers used various ways of speaking about this idea of checking on progress and sometimes explicitly made some sort of statement about attending to "deficiencies." Across four of the six districts, a total of seven teachers used phrases about "using data to drive our instruction" or "to help improve our instruction." Preparation for the state test most often meant finding out early where the student was likely to be on the state test so as to intervene, which necessarily overlapped with mastery of standards. Other purposes were identified less frequently (e.g., to place students in remedial or honors classes, to compare schools, or to evaluate teachers). When teachers gave idiosyncratic responses about their understanding of district intent, their answers tended to correspond to their own particular uses of the assessment information, which are explained further in subsequent sections.

As part of the larger study, Davidson and Frohbieter (2011) analyzed the consistency in understandings of purpose among district administrators, among teachers, and between administrators and teachers in the same district. In the Washington, Sinclair, and Taylor districts, over half of teachers mentioned consistent understandings of purpose related to mastery of standards, accountability, or preparation for the state test. In the other districts, less than half of teachers indicated common understandings of a district intent. For example, in Pittsfield SD, two of four teachers said that the purpose was to "ensure that standards are being taught" (Daisy) and to provide data about "areas that needed to have intervention for each student" (Jacob). But the other two teachers in Pittsfield described only reporting-to-the district purposes, to track how well they were following pacing guides and the like.

However, as noted by Davidson and Frohbieter (2011), even when teachers gave similar accounts of district intent their views might not match well with the reported aims of their district administrators. In contrast to teachers, district administrators frequently talked about school accountability and the need to monitor growth for subgroups of students. District leaders intended to direct effort toward the common curriculum that the assessment would either establish or enforce; administrators expected that an assessment aligned to the state standards and state test would provide a foundation for teacher collaboration and instructional improvement. Some district administrators expressed a desire to use the assessments as part of professional learning communities. For a couple of the administrators in the three districts with internally-developed benchmark assessments, there was the hope that teachers would learn more about students' conceptual understanding. Unfortunately, these more particularized goals were not mirrored in teachers' understandings of district intent; in some cases, principals and teachers

attributed to the district uses (e.g., placement of students) that were not mentioned by district-level administrators.

Professional Development

To understand the supports for access to assessment data and use of assessment results, teachers were asked, "What kind of technical support or professional development have you received to help you learn how to use this system?" For the most part, teachers received very little professional development (PD). Forty-six percent said, in fact, that they received little or no PD; seventy three percent said they received technical training on how to access data and understand scores. Only 35 percent mentioned any PD on "using" the data. Typically districts relied on a trainer-of-trainer model whereby an instructional coach or lead teacher was trained, who in turn would replicate the training in schools. Several respondents mentioned that training was provided for everyone in the first year of implementation but newcomers now had to rely on colleagues for help in accessing the system. In some schools, especially in Taylor SD, teachers continued to rely on the "testing person" to administer the tests and deal with testing related matters.

> We had in-services on interpreting reports on the website…knowing how to get reports, …knowing what the reports tell us ***(Washington SD, Alex)***.

> Our AP attended a half-day training then gave a one-hour seminar to teachers...We have half-hour discussions before and after administrations about how to access data and analyze results and understand how to modify teaching strategies [based on wrong answer choices] ***(Madison SD, Josh).***

> In August the testing person for our school goes over how to access scores and see how they compare to the district. Sometimes the testing person is available for one hour during in-services that teachers can opt to attend for help with analyzing results and breaking [results] down by standard ***(Taylor SD, Dolores).***

> We got less than an hour on how to upload to Data Director. No training is required to interpret results since they are straightforward ***(Pittsfield SD, Alyssa).***

Although some district leaders had aspired to the use of data through specific protocols (Davidson & Frohbieter, 2011), none of the teachers in the study mentioned receiving any training on how to collaborate. When grade-level teams or, in one case, professional learning communities were described they appeared to be idiosyncratic to the school or grade level rather than an organized and supported practice from the district level. For example, district administrators in Pittsfield and Sinclair mentioned the use of "inquiry" and professional learning communities, respectively, but three of four in Pittsfield and all of the teachers in Sinclair said

they got little or no PD. Some of the teachers in Sinclair referred to training in previous years with math modules to which the benchmarks were tied, but there were no longer meetings and discussions about kids explaining their thinking after the test became all multiple-choice. Ironically, at the school level, it was actually teachers in two other school districts who were more likely to mention getting together in "data teams" to review and respond to test results. In Burlington SD, three of four teachers said they got together with other teachers once per benchmark. "We look at everybody's data. We're open and we say, 'Wow, Meredith, you missed it here, but I got it, let me tell you, let me share with you'." Similarly, four of five teachers in Madison SD said they got together "to kind of get ideas of how to deliver it in a different way... when kids didn't get it" (Elizabeth). As we discuss in a later section, because of the nature of the test, these collaborative efforts in Madison SD were focused on individual test items rather than content standards.

Assessment Information

Mastery of standards, mastery of items. The mastery of standards and state testing themes, which teachers highlighted in talking about the *district intent* of the assessment, were also salient in how teachers described the information gained from assessment results:

> So then it gives me a report, like on this objective it was choose units of measure and do this and that. The people who didn't master it are these three people in that class, while the people who mastered it are all the rest of the class, so they did pretty good on these objectives. So I need to figure out what happened with these three people. You know, is it like they didn't understand it, or they're having a bad day, I've got to figure that out, especially if I see names—that I know them as good students, how come the good students missed it? ***(Burlington SD, Matthew).***

> Yeah, so what it does is it gives the result on what each child did in the different areas of algebra and functions, number sense, measurement and geometry, and statistics and data analysis. And then on the other side, you have a comparison, it has the class average, and then the school average, and then the district average, so you can compare to see how you're doing. And then they break it down by question, you know, what percent of the class chose what answer for each question ***(Sinclair SD, Rose).***

In fact, mastery of content was the primary type of information cited by over two-thirds of teachers, and was given as a secondary type of information by almost all of the remaining teachers. With only two exceptions, teachers brought class and individual student reports to their second interview and, when asked about the kind of information they got, pointed with a high degree of consistency to the results for each standard, for the class as a whole and student by student. In some districts, results were also reported for sub-standards and for each test item. In

one district, incorrect answer choices were also displayed. Standards might be as global as: Number Sense, Algebra, Data Analysis and Probability, Geometry, and Measurement. Whereas, a more tailored benchmark assessment might have detailed sub-standards such as: "Use the properties of complementary and supplementary angles and the sum of the angles of a triangle to solve problems involving an unknown angle," followed by the number correct out of five items representing these concepts.

Teachers differed as to whether they focused on "big picture" or more detailed information from test results. Excerpts in the Appendix section provide examples of responses at three levels of specificity: broad-progress information; standards-focused information combined with item-level information; and primarily item-level information. Although patterns by district were not perfectly consistent, there was a clear correspondence between the type of score report (and associated training) and the tendency for teachers to focus primarily on growth, standards, or items. For example, in the two districts using the NWEA MAP®, teachers described the broad information conveyed by the format of the assessment results:

> The information that is received is broken down into different areas of focus within mathematics, so it would be number sense or measurement and different things like that. So, you get a score for each section and also a cumulative score...and then throughout the year, it's mostly used just to show progress and achievement within students – from the beginning of the school year on, to make sure they're reaching the appropriate growth and they're going to be prepared to meet the state requirements when [the state test] comes around ***(Taylor SD, Margaret).***

> As specific as I can get is you can see patterns across the class and say, 'OK, I'll look and see. All these students are really struggling in their geometry or their operations.' I guess, that's about as specific as I can get...the thing is I haven't seen the actual tests that they take. I get their score back, but then I don't see the problems that they missed or whatever. It's just more of a score and then a category ***(Taylor SD, Heather)***.

Number sense was a commonly cited standard (in five of the seven districts) when teachers were asked for specific examples of information gained from assessment results and was often linked to a district priority or emphasis on the state test. Alex explained that because these standards were school-wide goals, teachers focused on them when receiving assessment results. Insights regarding the reason that a particular student was struggling with certain skills would have to be investigated through other assessments:

> I think the informal and the formal [assessments] that I do in my classroom, they serve a similar purpose in terms of guiding my teaching, as well as determining their grades and things like that. I think the NWEA is a much larger picture than those assessments give me. [My classroom] assessments give me a lot more about an individual student. The NWEA

> gives me more of a large class picture and like a placement of particular students, but that's still kind of in the large picture as opposed to that individual student really is having difficulties with integers. I don't know that particular piece of information from NWEA, but I do know that from my in-classroom stuff. I can know that the student has trouble with computation off NWEA, but I don't know that it's necessarily integers as opposed to fractions ***(Washington SD, Alex)***.

Heather (Taylor SD) similarly noted, "I just find that some of the other assessments that I give are more helpful than [the NWEA]." Even though two teachers, Margaret and Kelly, said that they looked at detailed information provided by the system based on individual students' scores, they too gained information from the test primarily at the level of broad, number sense and computational skill categories:

> *Margaret:* Some of [the scores were] shocking, because you wouldn't think an algebra student would be low on computational concepts and procedures. My other algebra student was low on number sense, which is also very strange. So knowing that, having a deeper focus within the classroom for that individual student and paying closer attention to that student's need is always something that's in my mind, in the back of my mind, when I'm teaching.
>
> *Interviewer:* Okay. And is there anything more you can say about what you learned about individual students or your class as whole from the MAP®?
>
> *Margaret:* No. It's somewhat vague. It's not as detailed as [the state test], so it's hard to really pinpoint. You still have to do a lot of guess and check in terms of where exactly their missing information lies ***(Taylor SD).***
>
> As far as the individual students – like, this top student right here, with having such a low number sense skill – 'cause [*sic*] they're only like in a 197. That's the lowest of the whole class. I know that they're going to struggle with the other ones. I know that if their number sense is low it's gonna [*sic*] affect every single standard that they're gonna [*sic*] have across the board, 'cause [*sic*] you have to have number sense to be able to do those other 7th grade concepts. And so, from that, like just seeing their individual scores really helps my planning, and knowing what I need to work on. So, this kid, I've actually worked with this kid specifically on number sense. Like, he gets extra work just doing the multiplying, the dividing, the fractions, the integers, like we really work on that, one-on-one ***(Washington SD, Kelly).***

As can be seen from the illustrative quotations in the Appendix section, when score reports included item-level information, some teachers continued to focus on standards-level information; others, particularly in Madison SD, attended to each item. Still some teachers remained focused on standards but appeared to go deeper with "item analysis," which suggests that they were attempting to understand which aspect of a standard was not being mastered. In

another section, we analyze the types of insights that teachers are able to gain from more detailed examination of score reports. Overall, we noted that they tended to focus on mastery of procedures and test-taking skills.

Evaluation of curriculum and teaching. In addition to describing the basic information gained from score reports, the majority of teachers also explained how they thought about the data and how they used it to reflect upon or evaluate their teaching.

> Then I can use that as a teacher to reflect on my own [and ask], 'OK, if they didn't make the gain, what am I doing that needs to change? How can I make sure all of my students are making the appropriate gains for a grade level?' ***(Taylor SD, Heather).***

> So when I look at this, I go this is not a reflection of my student; this is really a reflection of the teaching, not the student ***(Pittsfield SD, Daisy).***

> So I kind of look at it more as a benchmark for me, am I covering the things I need to cover, am I presenting them in ways that they'll be able to understand, am I presenting it in ways that's going to be familiar to them when it's presented in this kind of test format? ***(Sinclair SD, Rose).***

Teachers generally ascribed differences in performance to different amounts of time spent teaching particular standards or asked themselves what differences in teaching methods might have occurred. Poor performance might signal a need "to change instruction in some way;" yet, no specific examples were offered about how instruction might need to change.

Infrequent, procedural "insights." Because a goal of the study was to collect specific examples of insights teachers learned about their students from interim assessments, we were disappointed when teachers spoke only in general terms about the kinds of information provided. Therefore, in the final phase of analysis, we reread interview transcripts searching specifically for examples of substantive *insight*, defined as a) the specific aspects of topics or problems that students struggle with, b) examples of where understanding breaks down, or c) reasons for success or failure of mastery including difficulties with particular problem types such as ability to explain. Only 13 of the 30 teachers in the study gave any indication of insights derived from assessment results; in fact, most of the examples referred to either test-taking strategies or procedural errors. A few teachers examined multiple choice items with diagnostic distractors to better understand student misconceptions; others noted insights gained from grading constructed response items.

***Test-taking insights*.** In most cases, teachers' only insights across multiple districts were about test-taking skills rather than mathematics (e.g., "It helps me see that they need to learn to read directions better," "We take note on what kinds of accommodations the kids need," or "I'm

learning that they maybe went too fast, and so, it's just an opportunity for me to kind of touch base with them"). Janet (Adlington SD) noticed that she "had a large majority of students not do well with the word problem equations," and followed up by showing students a strategy for tackling word problems by "breaking down the equation and the solution." Similarly, another teacher recognized variations in student performance based on the question format:

> You know, I think it just is eye opening to see what kind of questions the kids struggle on as well, whether it be multiple choice or, you know, something where they have to write their thinking down. It's interesting to see what students and which students have trouble with specific types of questions ***(Madison SD, Elizabeth).***

The gist of other test-taking-skill insights tended to be aimed at anomalous aspects of the test-taking situation or format that kept students from demonstrating knowledge that they otherwise appeared to have in everyday classroom situations.

Procedural insights. More particularized information was noted by some teachers in districts with assessments that included diagnostic multiple choice items. Two of the five teachers in Madison SD described this type of report information:

> There were several different reports. They went into detail as to how many students got a particular question right or wrong; then it went into detail as to reasons why they might have gotten it right or wrong based on what they gave in an answer. You know, where did they miss the boat in terms of understanding the concept? What could we do differently in terms of our teaching strategies to enhance student learning for that – for any – actually, for any particular question – for every particular question? They had suggestions on how we could improve our teaching techniques and strategies... It's like, for instance, in 6th grade – factors and multiples. A student might answer with multiple rather than factor and then the test would have some information – this student did this because he, you know, he confused what a factor was with a multiple and here's how you can correct that ***(Madison SD, Josh)***.

> And then the teacher's editions are designed with, if he got "A," well, then he multiplied it by 15. It kinda [*sic*] tells you how they got that answer, if they didn't guess. It'll tell you, okay, well, they multiplied by 5, and then they were supposed to divide by 5 or something. And so the answers out of the test are designed so that if they make a mistake, they'll get one of the multiple choice, and so that just basically tells me what they did. ***(Madison SD, Molly).***

While noting diagnostic information here, Molly later stated that she did not use this aspect of the test to modify instruction because of the possibility of students guessing. Instead, she focused on item-level mastery and evidence of understanding shown in the constructed-response questions, as discussed below.

With the exception of Josh's comment above where he mentions the idea of conceptual understanding, insight responses were almost uniformly procedural, meaning that teachers pointed to computational errors or to algorithms that students had not yet mastered. In another district, Sophia pointed to procedural information gleaned from students' answers:

> So it tells me that the students chose an answer or the majority of this here because maybe they forgot to… they forgot that negative… that 2X minus 4X is a negative 2X and not just 2. So choosing that answer would be telling me that, so that printout tells me how many of the students chose a certain type of answer… And then if it was more of a random number, then that means they were all guessing and nobody really got that concept ***(Pittsfield SD, Sophia)***.

As with Madison SD's internally-created assessment, cited above, the diagnostic distractors on Pittsfield SD's textbook-based assessment reflected typical procedural errors. Keeping in mind that only 13 teachers gave specific examples that could be counted as insight, this tendency to focus more on procedures rather than students' conceptual understanding could be a function of the types of items available in particular assessments.

"Substantive" insights. We had difficulty locating substantive insights of the type we had originally hoped to find, where teachers had gained diagnostic information from the assessment that helped them pinpoint where a student's understanding was breaking down. One teacher in Sinclair found that students had trouble on the first trimester benchmark with requests to "explain, justify, or show your work" and then worked with students over the course of the year so that "students are getting much better at being able to write an intelligent statement that said what they did, or if it was work, to show reasonable, logical steps, that led to what we call 'path to solution.' That's a phrase that we use" (Sinclair SD, Patrick). This last example stands out because in combination with Patrick's subsequent teaching strategies, it appears to be focused primarily on students' mathematical understanding.

A few teachers in a district with an internally developed assessment that included constructed response questions noted insights gained from grading these items. While teachers in this district typically engaged in item-level analysis and review, they claimed that the constructed response portion of the assessment allowed them to better understand students' thought processes and conceptual understanding:

> I have one girl in particular who showed more understanding but she got the question wrong. Because she said, 'What's the least common denominator?' Here she said, 'Well it could be A because it could be one and a half over A.' Four and three-sixteenths. I mean that's higher-level thinking and she gets it wrong… And she did that on about three different questions. So I wouldn't know that about her if I just looked at the numbers ***(Madison SD, Michelle).***

> Yes, constructed response, and so based off of that, then I pull my small groups based on what I saw in their work, because there I actually have to see what they wrote and how they solved it versus these ***(Madison SD, Molly)***.

> Actually this student [who had shown 2+2=4 in her work]… she got a two out of three. This is how I would score it, because you actually have to square it, and that's 2x2, which is 4, so she got the right answer, she just didn't do the right concept. This would actually just be a little note, What if my radius was four, for example. Would you do the math the same way? [This] is what I would write on here ***(Madison SD, Elizabeth)***.

Elizabeth expanded on her use of constructed response answers in individual student conferences that further investigated students' understanding.

It is important to note that not all teachers described insights gained from grading constructed response items. For example, Daniel (Adlington SD) recounted his grading process by following a rubric based on the amount of work shown but only mentioned the information gained on mastery of standards by overall scores. Thus, despite our probing, insights were infrequent and largely based in test-taking or procedural skills. Rare examples of substantive insights were mainly derived from grading constructed-response questions, with some information gained from multiple-choice items with diagnostic distractors.

Assessment Uses

Reteach, review. In parallel to the most frequent assessment-information response being mastery of standards, the most frequent use of assessment information was to "reteach" or "review." Typically, teachers used the grid-like or graphic, standard-by-standard displays they received to identify the standards on which their class as a whole was weakest and designated these standards for reteaching. This predominant pattern varied systematically by district either because of the properties of the interim assessment or more likely because of differences in district practices and related professional development. In five districts, all 20 teachers mentioned reteaching or reviewing in describing how they used assessment information. By contrast, in the two districts using the NWEA MAP®, only one of 11 teachers described using assessment information to prompt reteaching. "I use it to drive what I'm gonna teach, 'cause all of my kids are having a hard time with integers….And I spend extra time on that and look at different ways of teaching it" (Washington SD, Kelly). As we describe in the next section, teachers in these two districts primarily reported using assessment results for grouping or placement of students in different class levels.

Within transcript data coded as *reteach, review*, we noted several variations. For example, all four of the teachers in Burlington SD used the term "standards" and talked about revisiting

the standards on which their students scored lowest. Although initially we thought that references to "concepts" by some teachers might refer to deeper insights about student thinking and conceptual understanding, we found after more complete analysis that standards and concepts appeared to be used interchangeably to refer to specific areas of content such as statistics, integers, or fractions. These examples are illustrative of the *reteach, review* category and reflect the majority of teachers who identify a standard, topic, or concept that needs to be repeated with the whole class:

> I could go back and look as a cluster at the whole class and see what standard was the one that was mastered the most by the majority of students. So that standard I feel like I could just visit pretty fast, and I don't have to take time explaining it again. But if it's a standard that most of my students failed to master, that's when I go back and reteach that standard ***(Burlington SD, Carol).***

> But after the [interim assessment], I really will focus on going back and trying at least to support them right away with what they didn't know. And then whatever it is, because it's basically the same thing, we go back and we review it, like fractions, fractions are another thing they forget ***(Burlington SD, Meredith).***

> Oh, definitely review it. Review specific math concepts that they're missing. Maybe it's a math concept or lesson that was a little bit more difficult than others ***(Adlington SD, Daniel).***

Consistent with our earlier descriptions of the procedural insights fostered by the type of item format used in Madison SD's benchmark assessment, teachers in this district tended to organize their reteaching around specific items and item formats:

> Number one is broken into grades, into periods, so I have first period and the percent of students that got it right, so I know that Question 1 on the interim assessment, only 12 percent got it right in first hour, meaning that's a whole group lesson I need to provide versus 72 percent got Question 7 right, which means I need to just do some small groups on a Question 7 example ***(Madison SD, Molly).***

> You know if there's quite a few that might have been in pink, then I might do a number talk with some of these questions. And that's a five to ten minute talk at the beginning of class ***(Madison SD, Melissa).***

> We kind of talk about each question, and then, we try to provide them opportunities via other quizzes or homework or in class work, where the questions are either composed the same way, where they all get practice as word problems or something that boggles them ***(Madison SD, Elizabeth).***

In moving from our analyses of assessment uses by category to characterizations of district patterns, we took care to ensure that each teacher was accurately typified by the selected

quotation. For example, Molly, Melissa, and Elizabeth (cited above) did not also provide other responses focused at the level of standards or general concepts. Thus, it is fair to note that reteaching in these cases was focused on individual test items and to suggest an association with the format of the district's benchmark reports.

Grouping, individual tutoring, and class placement. In addition to whole-class review on standards with the most prevalent weaknesses, over two-thirds of teachers in the sample also used interim assessment results to target interventions for individual students. Within their own classrooms, teachers most frequently used grouping to differentiate instruction. Following patterns noted previously, reteaching for subgroups of students might be organized around specific topics and skill areas or might target individual test items as illustrated by these two examples:

> From there, I will take the data and I look at, okay, if I have students who are low in number sense, then I have skills days in my classrooms. And so I'll group the kids together and I'll work on number sense with those students. If I need to work on algebraic reasoning, I'll work on that with a couple other students and really kinda break them up in my classes ***(Taylor SD, Cindy).***

> I choose a couple of these kids and match them with these so they can work together as a [study] group, so that they can reach each of them if I don't get around to all of them. And also, I reteach something. I could very well pull out these approaching ones, get them together and go over like four or five items out of this benchmark ***(Adlington SD, Mary).***

A few teachers offered individual attention through afterschool or Saturday tutoring, by assigning students to revisit a standard using computer software, or, in one case, by having low scoring students work with a volunteer.

Separate class placements—remedial, regular, and honors—were also seen as a means to address differences in student proficiency. All of the teachers in Washington SD and half of the teachers in both Burlington SD and Taylor SD reported using interim assessment results to place students in appropriate mathematics classes and sometimes to change a student's class level mid-year. For some of these teachers and schools, test results were used along with class performance to determine placement level. In some instances, however, test scores were the primary means for determining class placement.

> We do their schedules based on their NWEA scores, so that remediation class, those kids vary from a 195 to like a 210, where as this honors class generally varies from like a 235 to 250. And so, the kids are already kind of leveled and so within one class I don't have to differentiate too much ***(Washington SD, Kelly).***

> At the beginning of the year – like I said – most students are taking the test from the beginning of the year, so this is one way that we try to judge who is placed in the appropriate classes – especially those students who did not attend our school the previous year ***(Taylor SD, Margaret).***

> So based on these scores and the third trimester scores, along with their [state test], along with the grades that the teachers give, we'll probably put them either in the readiness or algebra. So that's what the benchmarks help us figure out ***(Adlington SD, Mary).***

We should note that the use of interim assessment data to track students into separate classes by ability level was surprising and was not mentioned as a policy intention by any of the district leaders. More likely, separately leveled classes were already customary in these schools and interim assessment results provided useful information to help categorize students by proficiency level.

Student feedback. Although not as frequent as other uses, 14 teachers reported using test results to give *feedback* to students. Unlike the literature on formative assessment, however, none of the descriptions or examples of feedback was about providing students with information about how to improve. Rather, in the context of interim assessment, feedback meant seeing which items you missed or learning to read the graphs so that you could see which standards you needed to work on.

> It has the information and then it graphs it so they can visually see where they are in comparison to the rest of the group... My goal is to get them to look at the specific weaknesses, so we actually take the graphs and look at, okay, here's numeracy where I'm weak. Here it specifically, it says, subtraction. Even down to an actual specific skill ***(Washington SD, Lisa).***

> I gave each of the kids a little piece of paper that had 1 through 20 on it, and it told 'em [*sic*] if they got No. 1 right or wrong, No. 2 right or wrong, No. 3 right or wrong. So each kid knew, if I said we're gonna [*sic*] go over No. 5, if they got it right or wrong. Not what their answer was, but just right or wrong. And so they know that they need to either pay a lot of attention to when I go over this question as a class, or if you kinda [*sic*] got it, you might wanna [*sic*] pay attention a little bit. And then after I've gone over the questions, I have them self-grade on that same piece of paper. They tell me whether they still need to work on it, if they need some more practice, they need some more teaching, or if they've got it ***(Madison SD, Molly).***

In some cases assessment feedback was more detailed and students could see how many more items they needed to get right in order to reach proficiency (Pittsfield SD, Sophia). In one school using the NWEA MAP® a pervasive focus on test-score improvement was described such that, "kids are expected to know their NWEA score" and to work toward their goal score on the

NWEA scale (Washington SD, Kelly). Similarly another district "used the interim assessments to move kids up or down on that board," so that coupled with their daily performance and their monthly performance in class "[they] [could] move a kid from partially to proficient or from unsat(isfactory) to partial or, unfortunately, the other way" (Madison SD, Josh). Or in this example, students are expected to understand test score results, identify the standards on which they are weakest, and work harder on those standards:

> So this particular student scored a 228, which is in the low range, and the projection would be partially proficient. And this is a common language in my classroom. I said, 'Well, where do we want you?' And the student will say, 'Well, we need to be in the proficient range.' So then we looked at the different sections, which this is broken into. It's almost – it's pretty much based on standards. So the standards that are the lowest are the ones that they'll need to focus on, which I have kind of marked for them ***(Taylor SD, Margaret).***

Note that, in this last example, Margaret is one of the teachers who provides students with computer modules to help them work on areas of weakness; yet, in her description of this conference with her student she does not offer any substantive feedback about conceptual errors or different approaches he might try to help him move forward. A general assumption underlying many of the responses appears to be that students will know what to do to improve once they receive feedback about problems missed.

Teachers also used assessment results to give feedback to students about test-taking skills. In parallel to assessment insights being about test-taking strategies, teachers reported giving feedback to students about not going too fast and writing more neatly. One teacher also commented that going over test results helped students take the test more seriously. "I have definitely learned how even just the little that I share this information with the student has helped them be more aware when they're taking the test that someone's actually looking at this and actually tracking what I do and making sure that I gain" ***(Taylor SD, Dolores)***.

Teacher Cases

During the interview process and again during the initial reading of transcripts, the interview team noted the prevalence of dissonant cases that were difficult to typify. A few teachers gave largely negative responses ("I don't think the information is valid, and I don't use it very much"), and a few gave entirely positive responses ("I gain very useful information and I use it regularly in my instruction"). However, most teachers were in the middle and expressed contradictions. For example, they might describe in a very positive way what useful information is provided but then state that they do not have time to look at it except for a few students or do not have time to reteach because they have to "move on" to keep up with pacing guides.

In the final stage of analysis, we developed two analytic tools to arrive at a more integrated summary of each teacher's assessment use. We gave each teacher a score from 1 (most negative) to 5 (most positive) indicating that teacher's overall judgment regarding the quality and usefulness of the assessment; moreover, we constructed distilled narrative summaries. Only four teachers (Meredith, Elizabeth, Kelly, and Sophia) in our sample of 30—all from different districts—received assessment-use scores of 4 or 5. Here we describe these more "fulsome" cases of interim assessment use and locate them in the context of their respective districts.

Meredith. Meredith teaches in Burlington SD, where she avows that the benchmark test drives planning and instruction: "I mean we teach to the test." She continually refers to meeting with other teachers to review data, plan, and possibly suggest new instructional strategies. Meredith especially notes the valuable guidance of a district coordinator that, upon request, assists teachers during structured teacher planning time to interpret benchmark test results and notice patterns:

> I'm going to go back, and I'm going to reteach. I am going to go back and I'm going to look at the stuff where the kids are just totally lost, [or] I'm going to take a look at where we just need to put a band-aid on it, you know, go back and say, 'oh, you guys got circumference and formula for area mixed up'—which they're still doing. And then we'll go back, we'll maybe get some more hands-on. I'm going to talk at my department meetings and say I need help, I'm not doing it right. The kids aren't getting it. Sometimes I'll even have one of my friend teachers come in and do a demonstration kind of lesson ***(Burlington SD, Meredith)***.

Meredith notes that sometimes her kids panic on the test even though they know it in class, but she also looks for common errors on individual test items and keeps track of consistent areas of difficulty from year to year such as ratios and proportions, volume of a cylinder, and fractions. Beyond a first review of the data to check for a certain level of mastery that "the district wants" for every student, Meredith uses the results to assess her own teaching:

> The other reports I get are how I did, as far as because to me, the way I look at it, if my kids all bombed it, then there's something wrong with me. I get information on each standard on how the class did, I get it from last year, I get it from this year, so I can do comparisons ***(Burlington SD, Meredith)***.

Although Meredith does not describe specific instructional responses targeted to students' misconceptions, she explains that reteaching is the main intervention strategy, which she does in warm ups:

> So I take that information back, and I try and fix it, I try and reteach. So what I'm really interested in, honestly, is where my bar is really low, on my bar graph, it's like, 'Wow, what happened here?' Because sometimes I really think I've checked for understanding, and I've

> done all the other things, and then I will go back at the text question too, to see, 'Let me look at this question and see is it me, or could there be something in the question that is really hard for the kids?' So I really use it to reteach ***(Burlington SD, Meredith).***

From this example and the report format, it seems that Meredith reteaches either standards or items, depending on the particular weaknesses shown in the data along with how well she believes to have taught the concept in question. However, she does mention a rare example of using insight gained from benchmark test results:

> Sometimes... it'll just be an error, like dividing fractions. They'll change the mixed numbers, but they'll forget to change the second number to a reciprocal and they'll multiply it out. So if I can find those simple errors, then we go back and say let's practice what we're doing here ***(Burlington SD, Meredith).***

In addition to her primary use of reteaching, Meredith mentions occasionally turning to more "hands-on" activities as well as having other teachers provide "demonstration lessons." Although she gives individual written feedback that indicates the standards that were both mastered and missed, it is up to students to take the initiative to ask for further help.

Meredith does note the lack of time; hence, her focus on the week following the assessment and reliance on a tutor to work with individual students. While she appreciates the information in the reports, she repeatedly states that there are "tons of data" that "take a long time" to get through. Likewise, she values team collaboration in data interpretation and use, but believes that teachers need about one day per quarter to achieve this, and that teachers would typically not want to miss being in the classroom for this purpose. She also notes the unfair comparisons of sixth grade students in middle and elementary schools that are using different tests and pacing as well as comparisons of students in honors in regular classes.

The other three teachers from Burlington SD report much more limited use of the district assessment, with negative comments becoming more salient. Eli likes the alignment with the state test but criticizes poorly worded questions and says there is not time to implement classroom changes. Matthew says he works with colleagues to identify missed standards for which there are likely to be a number of questions on the state test, but he repeatedly questions the usefulness of the information because he doesn't know how to help students who continually score poorly and have frequent absences. Carol gave highly abbreviated answers. She uses assessment results for planning, placement, and reteaching, but also emphasizes that there is not enough time to examine all of the individual results or to fully revisit missed concepts.

Elizabeth. Madison SD was the district where the greatest number of teachers reported gaining some kind of insight from the interim assessment, either from scoring the constructed

response items or from the types of wrong answers on multiple-choice items with diagnostic distractors. Three of the five teachers had an overall positive view of the assessment, with Elizabeth being the most positive. Elizabeth claims to learn valuable information about approximately 80% of her students; for various reasons, interim tests do not accurately gauge the knowledge and skills of the others. In line with Mandinach et al. (2008), she notes, "I have a hard time with saying what one form of assessment tells me about a student." Like Meredith, Elizabeth notes that kids become anxious at test time, but take it seriously because they know they are being compared to peers in the school and district. In fact, posters in the main school hallway display names of kids who are proficient.

Elizabeth emphasizes the insights she gains about both her students' test-taking strategies and specific misconceptions by grading the constructed response questions. She goes over the whole test for two days, and then reviews about one question per day on the three weakest standards or items for two weeks following the test. Elizabeth provides specific feedback and conferences with individual students on errors or misconceptions on the constructed response items; she does this for 20 minutes per day in the two to three weeks following the test. As Elizabeth notes, "if students score poorly they have to miss fun activities during numeracy time," and instead participate in small groups that focus on direct instruction of weak concepts.

Elizabeth speaks highly of the opportunity to collaborate with other teachers. She meets in weekly grade-level meetings, and consults with other teachers informally during common planning times. She notes that the teachers focus on item-level analysis of the three highest and weakest questions.

> We do that though in our community – or in our math meetings. We get together our data and we kind of say, 'How did your students do on – what was their weak points?' And we kind of find trends. And then, we plan together, either whole group or small group, interventions that we can do for those kids ***(Madison SD, Elizabeth).***

> I organize data, as far as – I itemize it by questions, and actually all of the other teachers I work with closely do it the same way, and then, we kind of compare across the board. How did our students do on number one? What was the trend we saw? What questions did they have the lowest percentage, and then, we talk about – or highest percentage. You know, what were we doing that was successful? ***(Madison SD, Elizabeth).***

Beyond the standard and item-level focus, Elizabeth claims that teachers take note of which students might need accommodations on the state test:

> We also take note on how many – like what kind of accommodations the kids need, because ...as long as you document all year long and are consistently doing the same thing throughout the year, prior to [the state] test, you can use that accommodation during the [state] test. And

> so, we definitely take note of who is taking longer and needs more time, who needs the test read to them, and these are students who don't necessarily have IEPs or are considered special ed. So, it's an opportunity for a regular student to kind of get those accommodations that they need in order to be successful ***(Madison SD, Elizabeth).***

Elizabeth notes that she spends about a week reviewing and interpreting the data on her own before giving feedback to kids and acting on the results. While she feels pressure to move on in order to cover the content on the next benchmark, she believes the time that she takes after each assessment is valuable in order to address weaknesses in anticipation of the state test.

Josh, another teacher in Madison SD, gives answers almost as positive as Elizabeth's. He especially likes the "what to do" guide that is quite useful in determining what to (re) teach if a student misses a question, and he reteaches regularly based on interim results. However, contrary to his principal's wishes, he does not use the district-generated model of "numeracy interviews" to follow up on individual student weaknesses. Molly also is generally positive. She is able to see what type of mistake students made on the test and directs her instruction to address those misconceptions. However, she notes that the test is not written very well and is often ahead of the pacing guide, and students are not motivated due to the extensive number of tests that they have to take. Two teachers in Madison SD are quite negative toward the benchmark assessment. Melissa said she didn't like the format and gave very limited responses about use of the information. Michelle doesn't use the results much and has very little faith in whether the assessment reflects the student's ability.

Kelly. Washington SD was one of the districts using the NWEA MAP® where all six teachers reported that test results were used for class placement and five of six said that they used it for grouping to allow for targeted instruction. Teachers in this district differed dramatically, however, in their judgments about the quality of the assessment and its use in their instruction. Kelly, who values NWEA MAP® the most highly, has already been quoted several times. She uses results extensively for planning and refers to a notebook of materials to link individual "RIT scores" to the district pacing guide and relevant instructional materials.

> [The] RIT score ranges then lead you into what you should be teaching the kids based off of that standard – so, off of computations, concepts and procedures, if their RIT score is between 231 and 240, I should be teaching them the stuff that's within this category to help move them forward into the next range[3]. [But] the biggest thing is the standard deviation. If it's below 15 we can teach the same stuff in the classroom. So, right here in this class I have

[3] See Figure 2.

two that are above 15, so I have to differentiate and teach two different levels within the RIT scores in my classroom ***(Washington SD, Kelly)***.

Subject: Mathematics
Goal Strand: Computation Concepts and Procedures
RIT Score Range: 231 - 240

Skills and Concepts to Enhance 221 - 230	Skills and Concepts to Develop 231 - 240	Skills and Concepts to Introduce 241 - 250
Conceptual Meanings for Operations	**Conceptual Meanings for Operations**	**Conceptual Meanings for Operations**
• Models algorithms using place value concepts (addition and subtraction with whole numbers)* • Models algorithms using place value concepts (multiplication and division with whole numbers)* • Uses a number line to determine the midpoint between a positive and negative number*	• Models algorithms using place value concepts (addition and subtraction with whole numbers)* • Models algorithms using place value concepts (multiplication and division with whole numbers)* • Uses models to multiply and divide fractions and connect the actions to algorithms* • Uses models to multiply and divide fractions and mixed fractions and connect the actions to algorithms*	• Uses a number line to determine the distance between a positive and negative number
Addition and Subtraction of Integers	**Addition and Subtraction of Integers**	**Addition and Subtraction of Integers**
• Adds integers with unlike signs • Adds several positive and negative integers • Solves real-world problems involving addition and subtraction of integers* • Solves problems involving addition and subtraction of integers*	• Adds integers with unlike signs • Adds several positive and negative integers • Subtracts integers* • Solves real-world problems involving addition and subtraction of integers (analysis)*	• Subtracts integers* • Solves real-world problems involving addition and subtraction of integers (analysis)*
Multiplication and Division of Integers	**Multiplication and Division of Integers**	**Multiplication and Division of Integers**
• Uses multiplication strategies to explain computation (e.g., doubles, 9-patterns, decomposing, partial products)* • Multiplies multiple-digit numbers • Divides a 4-digit number by a 2-digit number • Divides multiple-digit numbers • Divides numbers by powers of 10* • Solves complex word problems involving whole number division with remainder (e.g., 2-step, 2-digit divisor) • Uses division for multiple-step real-world problems (whole numbers)* • Solves real-world multiple-step problems involving whole numbers* • Multiplies integers with unlike signs* • Divides integers with unlike signs* • Solves real-world problems involving multiplication and division of integers*	• Divides multiple-digit numbers • Uses appropriate algorithms to represent multiplication or division with whole numbers* • Evaluates numerical expressions using the order of operations (whole numbers only) • Evaluates expressions using the order of operations, including exponents (whole numbers only) • Multiplies integers with like signs* • Divides integers with like signs* • Solves real-world problems involving multiplication and division of integers (analysis)* • Evaluates numerical expressions using the order of operations (using integers)*	• Evaluates expressions using the order of operations, including exponents (whole numbers only) • Solves real-world problems involving multiplication and division of integers (analysis)* • Evaluates numerical expressions using the order of operations (using integers)* • Evaluates expressions using the order of operations, including exponents (using integers)*

 CO 3.2.1
* Both data from test items and review by NWEA curriculum specialists are used to place learning continuum statements into appropriate RIT ranges.

Figure 2. Snapshot of score report that shows "RIT score ranges" to which Kelly refers.

Kelly notes that the standard deviation rule was decided at the school level, and describes a pervasive benchmark assessment culture. Teachers stress the importance of the test, and kids take it seriously because they know it will affect their "class schedule next year." Students know what their scores are as well as their learning goals in terms of reaching grade level proficiency or above.

> For some it's shown me that they don't do well with tests at all, and with others, sometimes it's amazing, because of the pressure that we put on our students about how they have to make grade-level and, with that knowledge, it sometimes pushes the kids a little bit harder. The scores are immediate, unlike [the state test], so they get their scores as soon as they hit "finish." So, it shows the kids right away what their – it's immediate feedback and that's huge for the kids ***(Washington SD, Kelly).***

After students receive their scores, they write personal reflections and discuss inconsistencies in their performance on specific standards in test and classroom settings. Kelly

notes that results could be used on a daily basis to "drive instruction," but is not clear on how often she employs the strategies that she mentions. In general, she looks for low levels of proficiency on particular standards at the classroom level, and responds by reteaching or grouping students by score:

> We can actually – we know exactly what benchmarks they're missing in certain standards, so if it's standard one and they're missing computation of adding fractions, they'll tell us that. When we pull up the information for their RIT scores we'll come across things that they need to work on this benchmark for this standard... [Then] I use it for [both heterogeneous and homogeneous] ability grouping ***(Washington SD, Kelly).***

Clearly, Kelly is conscientious about trying to find new ways of teaching when students do not do well. She sometimes responds directly to individual student scores on standards that give particularly useful insights, such as number sense. However, the remainder of her teaching strategies seemed to be focused on class-level results, with differentiation occurring on the rare occasion that some students were beyond the pre-determined acceptable variation within already leveled classes.

Other teachers in Washington SD are not so enthusiastic about the benchmark test. Lauren and Alex are neutral, saying that they use it primarily to get a "big picture" on student growth. Lauren appreciates the information that's there if she had more time to dig into it, but she takes results with a grain of salt because sometimes students drop 10 points even when she knows they are growing. Lisa is representative of the remaining three teachers who make little use of assessment results beyond placement and providing score reports to students. "I'll be really honest. As far as instructional purposes, I do not use the NWEA information that much."

Sophia. Our fourth exemplary case teaches in Pittsfield SD and had a hand in developing the quarterly assessments. Similarly to Elizabeth in Madison SD, the answer choices on the assessment let Sophia see, for example, whether a majority of kids are having a hard time with adding and subtracting negative integers. For the most part, she doesn't think the assessment tells her anything new (she already knows which students are low), but she uses it to determine what needs to be reviewed and especially focuses on standards that will be on the state test.

Students receive individual "proficiency reports" that give them feedback on their performance in each standard:

> The students look at their scores and identify which standards that they haven't mastered, and... if they're below... a 67 percent, they know... they need to focus on [that standard] when they're actually doing their assignments or studying for the tests... The printout gives the standards that were addressed on the test and it gives the students in the form of a bar, a percentage bar, and it tells them, like, say for algebra 4.0, it would have on that bar—every

> question that addresses algebra 4.0, however many they get correct, that addresses those standards. It would say, like, if there were four problems, they got two out of the four, it would mean 50 percent, so they know they need to solve one more in order to pass that standard ***(Pittsfield SD, Sophia)***.

In adjusting instruction, Sophia looks at class-level results at both the standard- and item-levels to identify "common mistakes." She then reteaches standards or items for which a majority of students scored low through warm-ups using a section called "spiral review" in the text book, while at the same time moving forward with the curriculum. She also uses a school-level strategy called "do now" that devotes 20 minutes to a quick review, as well as frequent problem-solving journal assignments. When results vary for students within a class, she differentiates to work directly with students who scored low and offer challenge problems to students who scored high. However, the decision to focus on weak standards is based upon their importance for the state test:

> What we need to look at for the [state test] is which standard is actually hit most, so which are the key ones and then which one could we go ahead and say, 'Okay, there's only one of that problem on the [state test],' so we put that one to the side and focus on the ones that have, like, four or five of them in the [state test] ***(Pittsfield SD, Sophia)***.

Like Elizabeth, Sophia notes that she cannot base her instruction entirely on the quarterly results, as she needs to move on in the curriculum and get other useful information from classroom assessments. However, she considers the results one piece of a broad picture that she holds of both her classes and individual students.

Other teachers in Pittsfield SD are not as engaged as Sophia in using the quarterlies. They reflect the themes already described of reteaching standards and especially emphasize preparation for the state test. Alyssa does test prep every Monday; Daisy mentions a computer program that provides practice problems. Jacob has little use for the quarterlies. He reteaches weak standards and offers after-school tutoring but criticizes the "factory model" of continually having to move on.

The other three districts lacked fulsome users of the benchmark or interim assessments, which meant that patterns of use ranged from solid, middle-of-the-road implementers to negative, recalcitrant users. In Taylor SD, the other district using the NWEA MAP®, all five teachers reported using assessment results for grouping. Margaret and Cindy look at assessment results when they plan instruction, but especially note vagueness of results and lack of integration with classroom assessments. Heather, Dolores and Rita make less use of the results because they do not have time or because they say they can learn more on homework or quizzes where they can see student work. In Adlington SD, all three teachers are middle-of-the-road

implementers, focusing on mastery of standards and using reteaching and grouping strategies. However, one teacher (Mary) emphasizes repeatedly that there is not enough time to look at or use the results in more detail. In Sinclair SD, Robert likes the benchmarks and uses assessment information for warm ups and afterschool help, but he says he does not gain particular insights about how to teach better and he concurs with two other, more negatively inclined teachers that the results come back too late to be useful.

Discussion

The purpose of this study was to gather data about how classroom teachers use interim and benchmark assessments. Specifically, we sought to examine how various forms of interim assessments influence teachers' judgments of students' understanding of mathematics as well as the influence of assessment results on teachers' instructional strategies. Although intended uses are relevant and clearly affect implementation, we wanted to move beyond rhetorical claims and gather examples of actual use and effects on classroom practice. We designed a two-stage interview strategy to make it possible to follow up on general claims – such as, "I use results to plan instruction" – in order to identify more specific examples of how assessment information was used in instruction.

Our work was informed by two separate but sometimes overlapping research literatures with distinct theories of action regarding the use of assessment information to improve teaching and learning. Formative assessment is closely tied to contemporary theories of learning and motivation and focuses on short-term adjustments to instruction, within the context of a specific lesson or unit of study, based on insights about students' understandings and/or misconceptions. By contrast, data-based decision making arose from efforts to apply total-quality-management ideas from business to the use of end-of-year summative tests in education and therefore implies data use resulting from longer-term periods of instruction. Not surprisingly, given longer time intervals and the close alignment between the formats of interim assessments and end-of-year summative tests, our findings show that interim assessment use more closely follows a model of data-based decision making rather than formative assessment. Nonetheless, given the ongoing confusion between interim assessment and formative assessment, it is helpful to continue to use both frameworks as points of reference when interpreting our findings.

Consistent with the accountability pressures that gave rise to them, the interim assessments in our study were used primarily to monitor student progress and to improve performance on state tests. The most pervasive positive effects appeared to be heightened attention to student achievement and intensification of effort. Whereas before teachers might have known that some students were struggling, they now had detailed lists of standards not met by individual students

as well as standards missed by the class as a whole. Typically teachers retaught the standards where students were furthest behind before moving on to the next topic. In one of the districts that provided item-level results, teachers spent class time reviewing the items that caused difficulties for the majority of students. Two-thirds of the teachers interviewed also organized class time or afterschool time to reteach standards with subgroups of students. These main patterns of use are quite similar to findings from Nabors Oláh, Lawrence, and Riggan (2010) in Philadelphia where teachers primarily used whole-group instruction during their reteaching week and less frequently used small-group instruction for less common benchmark errors.

Unfortunately, also consistent with research in Philadelphia (Bulkley, Christman, Goertz, & Lawrence, 2010), interim assessment data did not provide teachers with insights about what to do next other than reteach. As noted by Davidson and Frohbieter (2011), the overwhelming majority of administrators across districts hoped that interim assessments would be used for instructional purposes, but these aspirations were generally stated in quite broad and non-specific terms. Even when a few district-level administrators mentioned professional learning communities or data teams, there was no confirming evidence from teachers in these districts that they had received any guidance or professional development to this end. These findings echo the larger research literature on implementation of standards-based reforms. As noted by Elmore and Rothman (1999), test-based, incentive theories of change assumed that with sufficient pressure, teachers and principals would be motivated to find the means to improve instruction. Yet, their early findings from the 1994 Elementary and Secondary Education Act showed that many schools did not understand the changes that were needed and lacked the capacity to make them happen. More recently, in the context of No Child Left Behind, Carnoy, Elmore, and Siskin (2003) found that "better-situated schools" serving higher socioeconomic neighborhoods were more able to respond coherently to the demands of external accountability, identify the need for new curriculum content, and gain the necessary knowledge and skills. But most schools, especially those with the greatest needs, were not able to respond in this way, and there was little evidence that districts were able to provide the type of professional development needed to build capacity where it did not yet exist.

Across the seven districts in our study, professional development about how to use interim assessments was quite limited and primarily focused on accessing data on websites and reading score reports. In only two of the districts did there appear to be a pattern of collaboration, whereby teachers discussed assessment results and shared strategies for responding to areas of weakness; only in one of these instances did collaboration appear to be the result of a systematic investment by district officials. In the one case, the district was intent on making curricular changes based on the assessment and specifically set aside planning time to ensure attention to

results. In the other case, regular meetings appeared to be a continuation of math teacher meetings held years before to align the district benchmark with the state test.

Again these findings are consistent with lessons learned from the Philadelphia benchmark assessment case studies. In a rare case, when a committed instructional leader provided ongoing support, instructional resources, and a sense of moral purpose, Blanc, Christman, Liu, Mitchell, Travers, and Bulkley (2010) found that collaborations were coherent and sustained, and provided compelling evidence of what could be accomplished with data-driven systems. More prevalently though, without such leadership, Blanc et al. (2010) found that uses of interim assessment data were short-term and superficial, focusing primarily on predicting performance on the state test.

We note here that the research to date on data-based decision making and the use of interim assessments tends to be framed by organizational theory. As a result, the variables seen to have an effect on assessment use are organizational variables: leadership, time, alignment, professional development, teacher knowledge, and so forth. To be sure, our findings—regarding accountability pressures directing attention to the state test, limited professional development, and reteaching as the primary response—repeat these themes. Because of this focus, however, the quality of subject matter resources and the content validity of interim tests have scarcely been considered. Yet, our findings suggest that features of the assessments themselves may be shaping what teachers can learn from them. We began with questions about what teachers learned about their students from assessments, and probed for insights. We then reanalyzed the data looking for examples of substantive insight. Despite these efforts, we found few specifics about individual student learning, and when teachers did provide details, these "insights" were almost always about test-taking skills or step-by-step procedures for solving particular test items.

As one administrator explained, the hope is that interim assessments will help get beyond "blanket statements, students don't know fractions" and will instead help teachers "figure out: does a student understand the concept of part/whole and then is making operational mistakes or do they not understand the relationship in the first place" (Madison SD, PD Director). For this to be possible, Perie et al. (2009) argued that interim assessments should satisfy criteria such as the following:

> In general, to serve instructional purposes interim assessments intended to support diagnosis of students' understanding and misconceptions should include high quality open-ended tasks. All items, whether open ended or multiple choice, should be developed so that useful information about students' understanding and cognition can be gleaned from specific incorrect answers (p. 10).

The interim and benchmark assessments in the seven districts we studied did not meet these criteria. Madison SD came closest with its diagnostic distractors and two constructed response items; yet, due to the fact that the test items were not aimed at higher levels of cognitive demand, the information gained and subsequent teaching actions were not deeply conceptual. Teachers in Madison SD concurred with Perie et al.'s (2009) criteria when they pointed to constructed response items as the best means they had for learning about their students' thinking. Similarly, teachers in Sinclair SD recalled gaining greater insight in earlier years when more of the interim questions had been open-ended.

Other districts faced a substantive "Catch 22" in terms of their benchmark formats and hoped-for diagnostic insights. On the one hand, instruments that were very broad (reporting at the standards level only) provided little insight unless teachers followed up with further classroom based assessments. Conversely, more particularized reporting schemes that gave back item-level results tended to invite procedural insights and item-by-item reteaching. Note that this pattern was most likely caused by a combination of test features and lack of professional development to help teachers generalizing from specific items to larger conceptual strands underlying each standard.

Similar to findings from other recent studies, instructional uses of assessment results are tightly interwoven with attention to improving performance on the state test, raising proficiency rates, moving kids on the score scale, and so forth. As voiced by the four teachers with the most complete repertoires of assessment use along with many others, the interim assessments were part of a highly intentional, persistent set of efforts to raise test scores. Teachers described a benchmark assessment or accountability culture exemplified by posting students' scores in hallways and classrooms and giving feedback to students in terms of how many more items they needed to score correctly to reach proficiency. While these practices may have some positive effects on test scores following from the heightened attention to standards and sense of urgency, we should emphasize that the understandings of feedback as reported by teachers in this study were startlingly at odds with the recommendations regarding feedback that positively impact motivation and promote student learning in the formative assessment literature.

Dating from Sadler's seminal article (1989), critical ideas in conceiving how formative assessment was expected to work was that students would receive the type of feedback that would enable them to internalize the features of quality work and thus be able to monitor the quality of their own work "during the act of production itself" (p. 121). This meta-cognitive aspect of formative assessment is deeply substantive. Its importance is corroborated by repeated meta-analytic studies showing that feedback is most effective when it focuses on features of the task and provides students with specific, substantive advice about how to improve (Bangert-

Drowns et al., 1991; Kluger & DeNisi, 1996). Feedback is least effective, and may actually hinder learning, when it directs attention away from the task and towards the self, implying that the student is not an able learner. Grades and proficiency scores are in this second category of feedback. Thus, the hoped-for achievement gains expected to be leveraged by early and repeated assessment information need to be evaluated in light of the possible negative effects on learning and intrinsic motivation (Ryan & Deci, 2000) that are likely to occur when students are taught to measure their success in terms of test scores.

Enthusiasts for data-based decision making will see in our findings evidence that the majority of teachers use the results of interim assessments to make efficient use of the limited time they have available for reteaching and reviewing. Disappointingly, however, interim assessments do not provide teachers with information about students' thinking or diagnostic insights about their understanding that would suggest a particular way to intervene. Thus, the efficiencies appear to be largely managerial, rather than substantive. In responding to standard and item-level information, instructional improvements are often limited to student placement into leveled classes or ability groups rather than substantive feedback and attention to student misconceptions. While interim assessments can clearly be said to have focused attention on mastery of standards, a more definitive evaluation of their benefits will depend on independent verification of learning gains beyond practiced-for state tests. Given the limited, procedural information currently provided by interim assessments, as found in this and other studies, and the lack of information about what to do next, districts would be wise to consider more conceptual and open-ended assessment products and professional development strategies that are more directly linked to the learning goals and problem solving abilities that they aim to foster

References

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

Bennett, R. E. (2009). *A critical look at the meaning and basis of formative assessment*. Princeton, NJ: Educational Testing Service.

Black, P. J., & Wiliam, D. (1998a). Assessment and classroom learning, *Assessment in Education: Principles, Policy, and Practice, 5*(1), 7-74.

Black, P. J., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139-148.

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., Bulkley, K. E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education, 85*(2), 205–225.

Boudett, K.P., City, E.A., & Murnane, R.J. (Eds.). (2005). *Data wise: A step-by-step guide to using assessment results to improve teaching and learning*. Cambridge, MA: Harvard Education Press.

Bulkley, K. E., Christman, J. B., Goertz, M. E., & Lawrence, N. R. (2010). Building with benchmarks: The role of the district in Philadelphia's benchmark assessment system. *Peabody Journal of Education*, *85*(2), 186–204.

Bulkley, K.E., Nabors Oláh, L.N., & Blanc, S. (2010). Introduction to the special issue on "Benchmarks for Success? Interim assessments as a strategy for educational improvement." *Peabody Journal of Education, 85*(2), 115-124.

Burch, P. (2010). The bigger picture: Institutional perspectives on interim assessment technologies. *Peabody Journal of Education*, *85*(2), 147–162.

Carnoy, M., Elmore, R., & Siskin, L.S. (Eds.). (2003). *The new accountability: High schools and high-stakes testing.* New York, NY: Routledge Falmer.

Chappuis, S. (2005, August 10). Is formative assessment losing its meaning? *Education Week, 24*(44), 38.

Davidson, K.L. & Frohbieter, G. (2011). *District adoption and implementation of interim and benchmark assessments*. (CSE Technical Report). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Deming, W. E. (1986). *Out of the crisis*. Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Edmonds, R. (1979). Effective schools for the urban poor. *Educational Leadership, 17*(1), 20-24.

Elmore, R.F. & Rothman, R., Eds. (1999). *Testing, teaching, and learning: A guide for states and school districts*. Washington, DC: National Academy of Sciences - National Research Council.

Feldman, J. & Tung, R. (2001). Using data-based inquiry and decision making to improve instruction. *ERS Spectrum, 19*(3), 10-19.

Frohbieter, G., Greenwald, E., Stecher, B., & Schwartz, H. (2011). *Knowing and doing: What teachers learn from formative assessments and how they use the information*. (CSE Technical Report). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Halverson, R., Grigg, J., Prichett, R., & Thomas, C. (2005). The new instructional leadership: Creating data-driven instructional systems in school. *Journal of School Leadership, 17*, 159-194.

Herman, J. & Gribbons, B. (2001). *Lessons learned in using data to support school inquiry and continuous improvement: Final report to the Stuart Foundation*. (CSE Tech. Rep. No. 535). Los Angeles: University of California, Center for the Study of Evaluation (CSE).

Kluger, A.N. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254-284.

Lindquist, E. F. (Ed.). (1951). *Educational measurement*. Washington, DC: American Council on Education.

Mandinach, E.B., Honey, M., Light, D., & Brunner, C. (2008). A conceptual framework for data-driven decision making. In E.B. Mandinach & M. Honey (Eds.), *Data-driven school improvement: Linking data and learning*. (pp. 13-31). New York, NY: Teachers College Press.

Maxwell, J.A. (2005). *Qualitative research design: An interactive approach,* (2nd ed.). Thousand Oaks, CA: Sage.

McManus, S., Ed. (2008). *Attributes of effective formative assessment.* Washington, DC: Council of Chief State School Officers.

Nabors Oláh, L., Lawrence, N. R., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education, 85*(2), 226-245.

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.

Ryan, R. M. & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68-78.

Sadler, D.R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144.

Shepard, L., Hammerness, K., Darling-Hammond, L., Rust, F., Snowden, J.B., Gordon, E., Gutierrez, C., & Pacheco, A. (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teaching for a changing world: What teachers should learn and be able to do* (pp. 275-326). San Francisco, CA: Jossey-Bass.

Shepard, L.A. (2008). A brief history of accountability testing, 1965-2007. In K.E. Ryan & L.A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 25-46). New York, NY: Routledge.

Tyler, R. W. (1941). The functions of measurement in improving instruction. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 47-67). Washington, DC: American Council on Education.

Wayman, J.C. (2005). Involving teachers in data-driven decision making: Using computer data systems to support teacher inquiry and reflection. *Journal of Education for Students Placed at Risk, 10*(3), 295-308.

Wiliam, D. (2004, June). *Keeping learning on track: Integrating assessment with instruction*. Invited address to the 30th annual conference of the International Association for Educational Assessment, Philadelphia, PA.

Appendix

Examples of big picture growth and proficiency assessment information
"I get a breakout of an overall score that just kind of generally tells kind of like a grade level, but it's not supposed to be a grade level…and then I also have scores for every single standard, and that gives me a range for that particular kid, and then I know what areas the kid has a strength in and what they need help in…I learn in terms of their improvement. That's actually the biggest thing I'm looking for is just improvement, so I'm looking to see – you know, I can see their improvement" ***(Washington SD, Alex)***.
"But in my personal math classroom I don't really have much of an impact besides making sure that the kids are growing each time…Okay, so we get a printout that has the child's name on it, that has the score that they just got. And for math it has the breakdown by standard, so all six standards are on there. And then it has the range that they scored in for each standard and then the specific score. And so the range can tell you if it would be comparable to unsatisfactory, partially proficient, proficient, or advanced on our state test…" ***(Taylor SD, Dolores).***
Examples of mastery of standards and item-level assessment information
"Well, for the quarter assessments, this district quarter assessment that was done this year, I was one of the teachers that helped develop the quarter assessments because it's an accumulation of the standards that the students should have mastered. So we use it as a measurement of what the students have mastered, what we need to review, what we need to go over, and especially which ones are the main concerns that will be in the state test. …Classes, the student, and then we can do item analysis where we can look as a whole class, which one did they miss? Which one did they look like they were guessing on, almost everyone was guessing on or which was a common mistake?" ***(Pittsfield SD, Sophia)***.
(1) "They give us information like all the standards that have been mastered, and then we can look at it as a cluster, as a class itself, or we can get individual students to see what individual standards they have mastered, and which they need help on." (2) "So for example, the first student that I have, he mastered seven out of the nine standards we talked about. It gives me that, it gives me the percentage of proficiency, and it also gives me the questions that he got correct out of however many questions we tested on" ***(Burlington SD, Carol).***
(1) "Well, the test actually covers it, the state standards. So it'll take a few standards for the first test and break them down. And looking at our data, we can see which standard our whole class was low in, or which standard our class was high in, or averaging. And it helps us to focus on that particular area that they were low on." (2) "Well you can get a printout on each student, it tells you what they're having difficulties with, the different problems they're having difficulties with, the different questions. It gives you a whole breakdown" ***(Burlington SD, Eli).***
Examples of item-focused assessment information
"It'll give you percents of kids – how many percent – what percent got the first one right, what percent got the first one wrong. And so you can rank them based on 18 – question 18 was the most accurate; 4 was the next accurate. And so you can see which question they, as a class, did the worst and stuff like that" ***(Madison SD, Molly).***
"I don't know if I have anything that gives the school average. I think it tells each kid which problems were missed. Yeah, I'm pretty sure. Because I get them kind of mixed up in my head. We get some that do show which questions exactly were missed, and I am sure now that it's showing me which questions were missed, but I should be up on it, but I haven't used it that much, because it just…" ***(Sinclair SD, Patrick).***

Figure A1. Examples of big-picture growth, mastery of standards, and item-level assessment information.