CRESST REPORT 806

DISTRICT ADOPTION AND IMPLEMENTATION OF INTERIM AND BENCHMARK ASSESSMENTS

SEPTEMBER, 2011

Kristen L. Davidson

Greta Frohbieter

National Center for Research
on Evaluation, Standards, & Student Testing

UCLA | Graduate School of Education & Information Studies

District Adoption and Implementation of Interim and Benchmark Assessments

CRESST Report 806

Kristen L. Davidson and Greta Frohbieter
CRESST/ University of Colorado Boulder

September, 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: Davidson, K.L. & Frohbieter, G. (2011). *District adoption and implementation of interim and benchmark assessments*. (CRESST Report 806). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

DISTRICT ADOPTION AND IMPLEMENTATION OF INTERIM AND BENCHMARK ASSESSMENTS

Kristen L. Davidson and Greta Frohbieter
CRESST/University of Colorado Boulder

Abstract

As an outgrowth of the accountability requirements of the No Child Left Behind Act, districts are increasingly implementing "interim" or "benchmark" assessments. This report investigates various stakeholders' original purposes in adopting interim or benchmark assessments, ensuing implementation efforts, and actual assessment uses. We present findings from interviews with 24 district administrators and 14 principals in seven districts across two states and, where applicable, compare interview data from 30 teachers who participated in a larger study of classroom use of these assessments (Shepard, Davidson, & Bowman, 2011). District administrators often cited intentions that the assessments would be used to inform instruction. However, the realization of instructional purposes was limited by the type of information provided by predominantly multiple choice items, a lack of substantive professional development, and minimal coherence with respect to shared understandings of assessment purposes and uses across district, school, and classroom levels. Drawing from our results and other research, we provide recommendations for a successful interim or benchmark assessment system.

Introduction

Background

In response to a call to use data-based decision-making to improve student performance and achieve the proficiency goals of the No Child Left Behind Act, school districts have turned to assessments that provide more frequent information to teachers and administrators (Mandinach, Honey, Light, & Brunner, 2008; Young & Kim, 2010). Following this trend, "interim" or "benchmark" assessments that periodically test learning of recent content are proliferating nationwide (Bulkley, Nabors Oláh, & Blanc, 2010; Perie, Marion, & Gong, 2009).[1] While interim assessment systems claim to measure and improve student learning, the quality of information provided can vary substantially (Herman & Baker, 2005). To maximize the usefulness of the information, Li, Marion, Perie, and Gong (2010) and others emphasize the importance of district coherence with regard to the purposes for adoption of a particular system,

[1] Consistent with the definition of interim assessments provided by Perie et al. (2009), we use the terms "interim" and "benchmark" interchangeably throughout the paper.

efforts at implementation, and subsequent uses of assessment results. Yet, districts' processes to this end have been largely unexamined (Bulkley et al.; Mandinach et al.; Young & Kim).

The purpose of this report is to investigate the adoption and implementation of interim and benchmark assessment systems for middle school mathematics. We address the following questions:

1. What were stakeholders' reported *purposes and expectations* for interim assessments in their schools and districts?
2. What perceptions of the *implementation process* of interim assessments did stakeholders report, including assessment selection and professional development?
3. What actual *uses* of interim assessment results did stakeholders report? For example, were there improvements in instruction, course pacing, and/or collaboration? Were the reported uses *coherent,* i.e., aligned with the original purposes and expectations and with uses reported by other stakeholders?

We present findings from interviews with 24 district-level and 14 school-level administrators in seven districts across two states and, where applicable, compare interview data from teachers who participated in a larger study of classroom use of these assessments (see Shepard, Davidson, & Bowman, 2011).

Conceptual Framework

Advertisements for interim and benchmark assessments often cite documented achievement gains from the formative assessment literature in order to support claims that a particular system will improve student learning (Li et al., 2010; Perie et al., 2009; Popham, 2006; Shepard, 2005). However, it is important to recognize that the teaching and learning processes invoked by the *formative assessment* research base differ appreciably from the model of *data-based decision making* that characterizes the interim assessment movement. The literatures on interim and benchmark assessments and data-based decision making thus provided our framework for understanding districts' processes in meeting accountability requirements, ensuring alignment of content standards, and improving instruction.

Interim and Benchmark Assessments

We used the framework provided by Perie et al. (2009) to categorize the stated purposes for district adoption. The authors characterize interims as mid-range assessments that allow for the aggregation of results, and serve different purposes than both long-range summative assessments that gauge mastery of content and short-range formative assessments that inform daily

instruction[2]. The authors give twelve examples of reasons that districts adopt interim assessments, which they suggest can be understood through three classes of purposes: (1) *instructional* purposes intended to elucidate students' knowledge in order to respond instructionally in the current year; (2) *evaluative* purposes to monitor educational programs, curricula, or pedagogical methods in order to inform future changes; and (3) *predictive* purposes to anticipate performance on annual standardized tests or other measures.

Perie et al. (2009) emphasize the importance of clear and shared understandings of the purposes of an interim assessment during the adoption process. Specifically, the question, "What do we want the tests to tell us?" should determine the features of the most suitable assessment system (p. 9). In Appendix B, we summarize the authors' descriptions of these purposes, including each of their examples. Although the classes of purposes overlap (as the authors acknowledge), this framework provides a useful basis of comparison for our own findings.

Data-Based Decision Making

Rooted in Deming's (1986) business theory of "continuous improvement," an emphasis on data-based decision making characterized the education policy climate at the time of our study, and is closely tied to the proliferation of interim assessments. In the midst of accountability, testing, and standards-based movements, educators have been increasingly called upon to engage in an analogous model of *data-driven instruction* by documenting evidence of student learning and responding with instructional strategies toward improvement.

Virtually all data-based decision making frameworks emphasize the use of multiple sources of information, including tests, alternative assessments, classroom tasks, and the like (e.g., Boudett, City, & Murnane, 2005). The frameworks then outline organizational structures in which educators systematically interpret and act upon assessment results, such that coherence around the use of data exists both within and across district, school, and classroom levels. The authors often suggest that teachers collaborate toward common goals through "professional learning communities" (DuFour & Eaker, 1998) or through "data teams" that use protocols to identify a "learner-centered problem" to which instructional modifications are targeted (Boudett et al., p. 82). Finally, districts must develop the capacity for effective implementation by modeling these processes as well as providing schools with related resources, time, incentives, and professional development (Boudett et al.; Copland, Knapp, & Swinnerton, 2009; Halverson, 2010; Mandinach et al., 2008; Picciano, 2009; Wayman & Cho, 2009).

[2] Herman, Osmundson, & Dietel (2010) similarly define benchmark assessments as periodic evaluations of student learning toward long-term goals, with purposes and intended uses providing the basis for more specific features. They note that results may be used at the district, school, and classroom levels.

Some authors suggest that, while a coherent data-driven system is an improvement over disparate understandings and practices, it does not replace the need for attention to more substantive formative assessment (Boudett et al., 2005; Perie et al., 2009). Elmore and Rothman (1999) further note that professional development (PD) often falls far short, leaving teachers ill-equipped to fulfill the intentions of the system. They assert that standards-based reform can only be effective through an expanded theory of action that "make(s) explicit the link between standards, assessment, accountability, *instruction*, and learning" (p. 20, emphasis in original).

We use the data-based decision making lens to examine the coherence with which districts in our study adopted and implemented assessment systems. Drawing from Honig (2003) and Coburn and Talbert (2006), we present the distinct views of "top-level" and "frontline" administrators, and consider the extent to which seemingly different perspectives at district, school, and classroom levels might nonetheless reflect a similar theory of action.

Review of the Literature

As noted above, district processes with regard to interim assessment adoption and implementation remain largely uninvestigated. A review of the few relevant studies, however, reveals three general findings related to our research questions: (1) administrators and teachers hold different perspectives on the purposes, uses, and quality of an assessment system; (2) capacity building for effective implementation, especially in terms of PD, is often lacking; and (3) actual interim assessment uses may not necessarily reflect the district intent for its adoption.

First, adoption and implementation processes must attend to the fact that administrators and teachers typically have different perspectives on the purposes, uses, and overall quality of assessments. In examining perspectives on what counts as "valid evidence" of student learning and "appropriate use" of assessment results, Coburn and Talbert (2006) found that educators at different levels of authority rely on quite different criteria: district-level administrators want tests with sound psychometric properties; principals cite a “multiple measures” criterion for valid evidence; and classroom teachers tend to focus on measures that “capture student thinking or that are rooted in authentic instruction” (p. 485). Mandinach et al. (2008) likewise reviewed studies that contrasted the use of results by administrators, who monitor overall trends, and teachers, who rely on multiple sources in order to target individual learning needs. At the school level, Supovitz and Klein (2003) found that principals primarily used results to monitor progress toward school-wide goals, design PD, and determine school-wide interventions.

Second, with regard to capacity building and implementation, Young and Kim (2010) cite several studies that document the successes of instructional coaches, teacher collaboration (such as in professional learning communities), and allocated time for PD, data analysis, and

instructional experimentation. However, findings from the studies indicated that these practices are rare. For example, most principals did not receive needed training on how to guide teachers in transforming data into actionable knowledge. Similarly, Young and Kim (2010) point to the lack of, yet clear need for, pre-service and in-service training to develop teachers' "content knowledge, pedagogical understanding, and instructional skill" (p. 9) in using data effectively. In fact, the Means et al. (2009) national survey reported substantial proportions of teachers without the most basic knowledge for accessing and interpreting assessment results. Nonetheless, it is possible that districts and schools proceed through developmental stages of increasingly effective implementation (McLaughlin & Mitra, 2003).

Third, a lack of capacity building and shared understandings across educators at different levels can result in assessment uses that do not reflect the district's stated purposes. In a study that closely parallels our own, Blanc et al. (2010) found that instead of dialogue based in "pedagogical content knowledge" (Shulman, 1987) that would synthesize teachers' understanding of mathematical content and instructional strategies in order to respond to student needs, team meetings focused on targeting remediation for "bubble students" toward state test proficiency (pp. 212-213). In this way, the predictive purposes of interim assessments held the greatest weight, and the district intent did not cohere with the actual use of results. Young and Kim (2010) cite additional studies in Philadelphia, Milwaukee, and Wyoming that showed the ways in which incoherence among district educators' understandings with regard to the purposes, uses, and perceived validity of newly adopted assessments forestalled their effective implementation.

In parallel with the general trends in the research, our study examined districts' adoption and implementation processes, administrators' and teachers' perspectives, and district coherence.

Methods

This research represents a portion of a broader study on middle school mathematics assessment use in ten districts in California and Colorado. Five districts had only implemented interim or benchmark assessment systems, three had only partnered with collaboratives or universities to offer formative assessment strategies (largely through PD), and two had implemented both types of programs. This report therefore reflects the four districts in California and three districts in Colorado in which interim or benchmark assessments had been implemented. Pseudonyms are used for the districts and individual respondents.

Table 1 outlines the assessment systems and their general features. The Taylor and Washington districts used the NWEA MAP® computer adaptive assessment. Pittsfield School District (SD) created a quarterly assessment based on the sequencing in the *Holt* textbook series, and used online software to aggregate and report results. Burlington SD created an assessment

with teacher input on the benchmarks and sequencing, but contracted with Evans Newton for item creation and score reporting. All four of these assessments comprised exclusively multiple-choice items. The remaining three districts (Adlington, Madison, and Sinclair) provided internally-developed assessments, used online software for score reporting, and included one to four "enhanced multiple choice" (for which teachers gave partial credit based on work shown) or open-ended items (for which teachers gave scaled scores using a rubric), depending on the assessment. In addition, the multiple choice items on the Madison Interim Assessment were "diagnostic" in that incorrect answers pointed to students' possible misconceptions or procedural errors.

The research team conducted hour-long, in-person interviews with a total of 24 district-level administrators in the seven districts. Depending upon availability and personnel structure, we met with two to four administrators in each district, including the superintendent or deputy superintendent, the assessment director, a curriculum director (especially for mathematics), and the administrator responsible for PD. We asked the administrators about their districts' interim assessments as part of an interview exploring a variety of assessment types, for the broader research project. Administrators described the district's reasons and goals in adopting a particular system, methods of implementation (including PD), plans for continued use, assessment strengths and weaknesses, and current uses of results at the district, school, and classroom levels. Some of these questions were asked only to one or more particular types of administrators; for example, PD coordinators were not asked about assessment selection, as they were not likely involved in it. Administrators were also able to opt out of answering sets of questions on topics with which they were unfamiliar. All administrators were asked about the district's intentions in adopting an interim assessment, so that we could examine coherence in this area. The protocol used for our administrator interviews is included in Appendix A.

District-level administrators recommended two schools in which they felt the assessment system was well implemented. We proceeded to conduct hour-long, in-person interviews with the administrators at each of these schools, for a total of 13 principals and one assistant principal.[3] Interview questions for the school administrators focused on implementation processes and uses of results rather than the purposes of the assessment. Principals recommended one to three teachers at each school with exemplary assessment practices, with whom we conducted in-depth, two-stage interviews.

[3] In one of the schools in Sinclair SD, the Assistant Principal was primarily responsible for management of the benchmark assessment system.

All interviews were audio-recorded and transcribed. The authors of this paper coded the district-level and principal interviews using matrices of "organizational" (based in the conceptual framework and study design) and "substantive" (emerging from the data) categories (Maxwell, 2005, p. 97; Miles & Huberman, 1994), documenting both confirming and disconfirming evidence for each case (Erickson, 1999). Seven organizational categories were derived from the interview structure, including: Reasons for Adoption; Implementation; District-Level Uses; School-Level Uses; Classroom-Level Uses; Strengths; and Weaknesses. Responses within each category were coded using the Perie et al. (2009) scheme outlined in Appendix B, as well as substantive codes that were created from the interviewees' responses, as listed in Appendix C.

Table 1

Interim and Benchmark Assessment Systems and Features

District	Assessment	Features of exam	Frequency	Features of score reports
Adlington	Adlington Benchmark Assessment[a],[b]	25 MC[c] + 1 CR[d] (Teacher-graded with district rubric)	Trimester; final test cumulative	Class, subgroup, & student-level proficiency for standards & substandards; Lists substandards to focus on based on items missed
Burlington	Burlington District Assessment	25-65 MC (Items/scoring by Evans Newton)	Quarterly; not cumulative	Class, subgroup, & student-level proficiency levels & scores for standards & substandards; Class-level item analysis; Grouping report of "mastered" & "non-mastered" students for each objective
Madison	Madison Interim Assessment	18 MC + 2 CR (Teacher-graded with team-created rubric; some items based on Connected Math)	Trimester; not cumulative	Class, subgroup, & student-level scores & responses by standard & item
Pittsfield	Pittsfield Quarterly Assessment	25-50 MC (Based on Holt text)	Quarterly; cumulative	Class & student-level scores for standards & substandards; Class- & student-level item analysis
Sinclair	Sinclair Benchmark Assessmentb	16 MC + 4 "Enhanced MC" (Teacher-graded)	Trimester; not cumulative	Class & student-level scores for standards; Class & student-level item analysis
Taylor and Washington	NWEA MAP®	CAT[e]; 50-52 MC	Trimester; not cumulative	Class & student-level RIT[f] scores for standards; DesCartes tool lists skills & concepts to focus on by standard & RIT score range

Note. [a]Assessments with school district names were internally developed.
[b]Adlington and Sinclair school districts also used the POWERSOURCE© formative assessment strategy.
[c]MC = multiple choice; [d]CR = constructed response; [e]CAT=computer adaptive test; [f]RIT=Rasch unit.
Scores are based on national normative comparisons for grade-level performance.

Findings

We present findings across all districts for each research question: 1) the stakeholders' *purposes* and *expectations* for the interim or benchmark assessments; 2) the stakeholders' *perceptions* of the *implementation processes*, including assessment selection and professional development; and 3) stakeholders' *perceptions* of the *actual uses* of the *assessment results* as well as the *coherence* within each district with regard to *original purposes* and *expectations*.

Research Question 1

What were stakeholders' reported purposes and expectations for interim assessments in their schools and districts?

When asked, "What were the district's goals for the assessment system?" administrators typically addressed district intent as understood at the time of adoption as well as current goals for using the system.[4] To gauge the extent to which these purposes were understood by the primary users of the assessment, we also asked 26 teachers in six of the seven districts, "Can you tell me what the district's expectations were in deciding to use the assessment system?"[5] Here we present respondents' views using Perie et al.'s (2009) instructional, evaluative, and predictive purpose categories. We include the views of teachers and only district-level, or "central office" administrators, as principals were not asked about assessment purposes. We note variation in perspectives held by five "top level" administrators (including superintendents and deputy superintendents) and 19 "frontline" administrators (including three PD directors, six assessment specialists, and 10 curriculum specialists) (Coburn & Talbert, 2006; Honig 2003). Because of substantive differences between frontline administrators' responses in districts with only multiple-choice exams, as opposed to those including some constructed response items, we distinguish findings for this group of respondents by assessment type.[6]

[4] Not all administrators were employed in their position during district adoption of the assessment, such that views of original district intent and current goals were indistinguishable. This was particularly the case in Washington SD, which was in its eighth year of using the NWEA MAP®.

[5] Teachers in Adlington SD only discussed the POWERSOURCE® formative assessment strategy.

[6] For simplicity, we use the term "constructed-response" to refer to both open-ended and "enhanced multiple-choice" items.

Table 2

District-Level Administrators' and Teachers' Understandings of Interim Assessment Purposes

Purpose	"Top-level" administrators *(N=5)*	"Frontline" administrators		Teachers (*N*=26)
		MC only (*N*=11)	MC[a] + CR[b] (*N*=8)	
Instructional	80%	82%	100%	54%
Broadly "inform instruction"	80%	82%	63%	31%
Gauge content mastery	0%	18%	75%	35%
Provide insight into conceptual understanding	0%	0%	63%	0%
Provide feedback to teachers and/or students	20%	18%	25%	12%
Evaluative	100%	91%	100%	38%
Aggregate achievement data	100%	55%	63%	27%
Standardize curriculum	80%	45%	88%	0%
Monitor coverage and pacing	40%	9%	50%	19%
Evaluate instruction, curriculum, and/or pedagogy	40%	0%	13%	8%
Predictive	40%	45%	25%	27%
Predict and track progress toward state test performance	40%	36%	25%	12%
Provide feedback to improve state test performance	20%	18%	13%	19%

Note. [a]MC=multiple choice; [b]CR=constructed response

Overview. About 80% of all district-level administrators cited both instructional *and* evaluative purposes for assessment adoption. Half of district leaders emphasized "informing instruction" as the primary intent, while describing evaluative goals - such as aggregating achievement data, standardizing the district curriculum, and ensuring appropriate pacing - as of secondary importance. Only a few district administrators naming both goals described evaluative aims as primary, with the remainder giving fairly equal attention to both aims. Several teachers recognized evaluative aims, but the majority understood the district intent as instructional in nature, especially with regard to ensuring students' mastery of the standards. Finally, only one or two central office administrators in each district cited predictive purposes of preparing for the state test, indicating a lack of coherence about this aim at the district level as well as a stronger

interest in instructional and evaluative aims. Some teachers, however, perceived a singular district intent of predicting and improving state test performance. It should be noted that most teachers expressed uncertainty about the district's intent for the assessment, expressing their views on this topic in a largely speculative manner.

District-level administrators' and teachers' understandings of the purposes of the interim assessments used in their districts are summarized in Table 2. The Perie et al. (2009) framework shown in Appendix B is condensed here. Overall percentages given for instructional, evaluative, and predictive aims reflect the proportion of respondents that cited any example within those categories.

Instructional Purposes. District-level administrators in five of the seven districts emphasized an intent for interim assessment results to inform instruction. In the other two districts, Adlington focused on the summative information provided by the tests, and Washington focused on monitoring school accountability and predicting the state test. A goal of promoting teacher collaboration around using assessment results to inform instruction was likewise mentioned frequently, but only the Burlington and Madison districts showed evidence of this practice occurring consistently in the schools.

District-level administrators noted instructional aims as an important part of the theory of action driving interim assessment adoption. For example:

> Certainly the whole interim assessment movement really revolves around the theory of teaching and learning and planning. Doing some teaching, looking at data to determine …where are kids and then adjusting. Part of all this is to get some standardized ways to get our teachers to use [data] in this teaching and learning cycle ***(Taylor SD, Assessment Director)***.

Parallel to the evaluative purpose of standardizing the curriculum (discussed below), instruction was expected to respond to class- and student-level information on proficiency in the standards.

While top-level administrators gave broad references to "informing instruction," frontline administrators mentioned more specific goals of gauging content mastery and providing feedback. The majority of frontline administrators in districts with constructed response items also explained a goal of providing insights into students' conceptual understanding, and often explicitly connected this aim with the decision to include constructed response items:

> We couldn't let go of getting inside a kid's head and knowing. So we put... no more than [four] times where we ask [kids] [to] explain your reasoning, or show your work, or justify your reasoning or whatever. So that a teacher would have at least a few instances of [how]

> [kids] [are] retaining learning over time, where they could get into the kid's head and hear in the kid's words that piece about, so, what does slope mean in this context ***(Sinclair SD, Curriculum Director)***.

Other district leaders suggested a difference between interpretations of assessment results based on broad content areas and those that provide more specific insights:

> [The] [interims] are to be used for planning... we experience teachers making a blanket statement; students don't know fractions, okay? And what the assessment allows us to do is figure out - does a student understand the concept of part/whole and then is making operational mistakes or do they not understand that relationship in the first place... it's looking at the conceptual understanding... Then we can correct minor mistakes more easily and not teach a unit on fractions because students make mistakes in solving some problems using fractions ***(Madison SD, PD Director).***

The Assessment Director likewise noted that the Madison Interim Assessment included items from the *Connected Math* textbook because they were "conceptually based."

While teachers most frequently understood district intent in terms of instructional purposes, they gave only vague descriptions of informing instruction based on periodic gauges of content mastery:

> 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… ***(Sinclair SD, Rose)***.

In fact, no teachers referenced a district intent related to conceptual understanding. A few teachers did report gaining insights through constructed response items when describing their actual practices (see Shepard et al., 2011). However, the grading of these items was typically optional and difficult to complete in the limited turnaround time before assessments were due to the district. Coupled with the emphasis on aggregated reports based only on the multiple choice items, it is understandable that teachers did not perceive the constructed response items to be a high priority for their districts.

A few teachers understood the district intent as encouraging a particular practice such as differentiating instruction, grouping, or placement of students in leveled classes:

> We are expected to use the information to differentiate instruction within our classrooms, so using it to group kids, as well as to know the levels that they are at to actually meet them at that level and take them to the levels above that ***(Washington SD, Alex).***

> I think there were a lot of intended reasons for using MAP®, but I know that the biggest one was class placement.... So when a student comes in new or at the beginning of the school year, if they meet a certain standard on the math portion of the test or the reading portion of the test it determines whether they get placed in a grade level class or an advanced class, an honors class, an AP class ***(Taylor SD, Dolores)***.

While administrators did not mention these specific instructional responses as district goals, they did sometimes acknowledge them as classroom-level *uses*, as reported below.

Evaluative Purposes. First, district-level administrators commonly cited an interest in aggregating achievement data such that schools, classrooms, and subgroups of students could be compared. Because the aggregation of results is central to Perie et al.'s (2009) definition of interim assessment, it is not surprising that districts would aim to benefit from this type of information. However, all districts recognized that in order for aggregated results to be useful - either for programmatic or instructional purposes - a common curriculum was essential.

In fact, we found that district leaders' reasons and goals for assessment adoption were strongly influenced by the extent to which a common curriculum had been established in the district. Perie et al. (2009) note that "(‘district leaders') goals may be to enforce some minimal quality through standardization of curriculum and pacing guides..." (p. 8). Districts without a common curriculum thus sought to establish one, while districts with an established curriculum aimed to ensure consistency of coverage and appropriate pacing.

Because the Burlington, Pittsfield, and Taylor districts did not yet have common curricula, administrators in these districts saw interim assessment as an opportunity to focus instruction on the state standards. As the Curriculum Director in Taylor SD noted, "The standards we have for kids to learn are really defined by the assessments they use." The Deputy Superintendent in Pittsfield SD likewise recounted interim assessment adoption as a means to standardize the curriculum:

> The way we got to where we are is that we initially began to work with the whole idea of curriculum and instruction and assessments and monitoring. Our curriculum was not truly written down in a way that was very useful... There was no agreement on what standards we were addressing... So trying to get everybody to pull this all together, we went with a company who could—who would provide leverage to do that for us... They had some very old material and the folks who were here could talk about standards, but there was nothing that you could actually provide, that everyone had agreed to and everyone was teaching to. So this was a company that would provide us the leverage because in order to do the assessments and analyze them you had to have the curriculum online... So it forced us to actually give it them ***(Pittsfield SD, Deputy Superintendent).***

District leaders in Washington SD expressed concern that the NWEA MAP® assessment did not align with state standards; they were in the process of developing a standards-based curriculum and seeking a corresponding benchmark assessment.

Interestingly, each of the three districts that included constructed response items on their assessments (Adlington, Madison, and Sinclair) had already established a common curriculum aligned to state standards. For these districts, evaluative goals for the assessments centered on ensuring appropriate pacing and consistency, increasing teachers' awareness of the standards ("unpacking the standards"), and encouraging a "common conversation." The Adlington Curriculum Director noted that due to the assessment, "our teachers become aware of the standards and understand the expectation of the grade level better." She also stated:

> First of all, we can monitor the coverage [and] pacing. We have situations …where the teacher only covered 8 chapters or [less] out of the 12 chapters. Second is to see if the student developed proficiency during that trimester, so you don't wait until the end of the year or the standardized test.

In addition to state test preparation, administrators cited concerns that teacher variation in curriculum coverage resulted in students' inadequate preparation for subsequent coursework.

It is important to note that an evaluative purpose of standardizing the curriculum was often seen as a necessary step to achieving instructional aims. As the Curriculum Director in Taylor SD explained:

> ...we're a very site-based district and we know that we need to have our curriculum and assessments in alignment so that, that can really help get to improved instruction in the classroom. So I think the... big reason is to align curriculum and instruction in our district.

In this way, standardization of the curriculum intersected with an instructional purpose of gauging proficiency in the content, which aligned with teachers' understanding of the district intent to ensure that all students "have *mastered* the standards" (Burlington SD, Carol).

Some teachers also noted an intent to monitor pacing and curriculum coverage. This evaluative aim affected teachers' planning:

> One [reason] is to help drive instruction. And the other is to ensure that standards are being taught. And so that people or teachers are focused on what it is that they're supposed to teach, where they should be, where they should end up, what the student needs to learn ***(Pittsfield SD, Daisy)***.

Because administrators' intent to ensure *coverage* of standards was understandably viewed by teachers as a means to measure whether students *mastered* the standards, it was difficult to

disentangle evaluative and instructional aims. Even the two districts that did not emphasize instructional aims appeared to be moving in that direction, as Adlington included constructed response items to inform instruction, and Washington was redesigning the curriculum and seeking a better aligned assessment. Of course, standardization of the curriculum was also clearly tied to the standards reflected on the state test.

Predictive Purposes. While they were the least frequently cited overall, *predictive* aims were mentioned by at least one respondent in six districts. Given the frequently cited goal of standardizing the curriculum to focus on state standards (and thus with the state test based on these standards), predictive purposes may have provided an additional push to adopt an interim assessment system. Some administrators further noted the importance of specifically aligning the assessment content with the state test for predictive purposes:

> ... the only information we have as a district is [the] [state] [test]... Our teachers and our principals, our principals particularly said, "Isn't there some way that we can be working along the year that can be predictive of how well our kids are going to do in [the] [spring]?... Aren't there some [assessments] that could help us along the way, before we get to the big, huge, and then boom! And it's high stakes now. For every school, major high stakes. So, part of the reason that they are mandating [the benchmark assessments] is that we know that [they] are standards based ***(Sinclair SD, Math Director)***.

In Taylor SD, there was general agreement among both administrators and teachers that preparation for the state test was central. Teachers' responses echoed that of the Assistant Superintendent:

> I know that's a common selling point, a little slogan, but it actually is pretty well aligned with [state] standards and it makes predictions overall that are pretty good between where you score on MAP® and how you do on [the] [state] [test]. So some of that is helpful to people too, not so much to raise [state] [test] scores; more to say where are kids in learning this stuff ***(Taylor SD, Assistant Superintendent)***.

> My impression …is that it was an assessment created to help teachers get an idea of how students are going to be performing on the …[state] [test] because it was a faster and easier way to get data and we could give it to a student multiple times throughout the year and analyze in which specific areas students are struggling so that we can help them with that for the [state] [test] ***(Taylor SD, Margaret).***

Although less commonly reported in other districts, predicting and preparing for the state test was the *only* goal of the assessment cited by five additional teachers in four other districts:

> From my knowledge, they were using it to see if students were growing at the level they should to succeed on [the] [state] [test]... The questions were tied to [the] [state] [test]… and so if they can't do those, we need to go back and reteach ***(Madison SD, Molly).***

Although it is considered an evaluative aim, we found that standardization of curriculum linked the three purposes named by Perie et al. (2009) by providing a common base from which to inform instruction, ensure coverage of standards, and at the same time prepare for the annual accountability measure.

Research Question 2

What perceptions of the implementation process of interim assessments did stakeholders report, including assessment selection and professional development?

We asked district-level administrators and principals to describe their assessment implementation processes, as well as specific efforts to provide related PD. We likewise asked teachers to comment on the PD that they had received, including whether it focused on "how to use the system or how to use the results."

Selection and Implementation Processes. It is interesting that along with the above intentions for the assessments they adopted, all but one district-level administrator emphasized the importance of choosing an assessment system that would ensure the smoothest implementation process. In the Taylor and Washington districts, administrators selected the computer adaptive version of the NWEA MAP® assessment - despite a lack of needed technological infrastructure - largely because of familiarity with the previous paper-and-pencil version. The Madison and Pittsfield districts used items from their respective text series (*Connected Math* and *Holt*) in their assessments in part because of familiarity. The other districts chose to develop their own assessments, as they were accustomed to internal decision-making by committees that included teachers and instructional leaders. A few administrators mentioned the need to get an assessment in place as soon as possible and to avoid politically-charged delays in its implementation. In this way, the extent to which the format of the testing system aligned with current district practices was salient.

After selecting an assessment system, most districts proceeded with a "trainer-of-trainer" model in which the district trained instructional coaches or lead teachers who were then expected to lead PD efforts in schools and provide teacher support. In practice, the extent and quality of the replication of this training varied by district and school. The Taylor and Washington districts primarily trained test proctors to administer the exams in school computer labs, which resulted in many teachers being largely disconnected from the assessment process. In the other school districts, the trainer-of-trainer model was used to provide PD to teachers on using data to inform instruction by way of teacher collaboration, curriculum planning based in content knowledge, goal setting, and specific instructional strategies.

Even though central administrators claimed that principals and teachers were included in the selection and implementation process, principals in five districts stated that compliance with district direction to use the system was simply mandated. In some cases, lead teachers were involved in determining curriculum sequencing and selecting appropriate test items from the texts. However, most teachers' participation consisted solely of reporting poorly written or biased items to the district.

Professional Development. To varying degrees, respondents at every level described PD that trained teachers on two different topics: using data to inform instruction, and accessing and understanding assessment results ("technical training"). Table 3 quantifies the reports of PD by district-level administrators, principals, and teachers.[7] Administrators most frequently cited efforts to provide PD on using data to inform instruction, especially through encouragement of data-driven dialogue in team meetings. However, the great majority of teachers claimed to have received only technical training or very little PD related to the assessment. While teachers in two districts discussed participation in team meetings, none mentioned receiving PD that would aid in this process. Instead, the teachers who did describe PD on informing instruction cited a focus on instructional strategies, such as differentiation and grouping techniques.

Table 3

Reports of Professional Development Offered and Received

Type of professional development	District-level administrators (*N*=24)	Principals (*N*=14)	Teachers (*N*=26)
Use data to inform instruction[a]	75%	43%	35%
Encourage data-driven dialogue in team meetings	42%	43%	0%
Guide curriculum planning and increase content knowledge	25%	29%	8%
Set goals for academic growth for individuals or sub-groups	17%	29%	4%
Provide specific instructional strategies	21%	21%	23%
Technical training	29%	29%	73%
Little/ no PD	-	-	46%

Note. Percentages do not add to 100% due to overlap in types of implementation cited.
[a] Percentages for "Use Data to Inform Instruction" represent any mention of the four types listed.

[7] Because reports were similar for top-level and frontline administrators, and PD was more specific to districts than to assessment format, we simply present overall percentages for respondents at each level.

Use Data to Inform Instruction. Respondents' descriptions of staff development on using assessment results to inform instruction comprised the following categories: (1) encouraging data-driven dialogue among teachers; (2) guiding curriculum planning and increasing teacher content knowledge; (3) setting student goals with an emphasis on targeting disparities in achievement among sub-groups; and (4) providing specific instructional strategies. Staff development was centered on standard- and item-level information, with only one respondent describing PD on the analysis of constructed response items.

Encourage Data-Driven Dialogue. In line with the literature on data-based decision making, administrators in five districts discussed efforts to promote data-driven dialogues in content area or grade-level team meetings. Respondents in the Burlington and Taylor districts talked in general terms about encouraging teacher collaboration in "data teams," while those in other districts named the use of specific protocols: (1) Madison SD purchased a protocol from a prominent assessment company; (2) Pittsfield SD developed an "inquiry-based" model with an outside consultant; and (3) Sinclair SD created a "professional learning community" (PLC) protocol. The protocols were generally based on standard- and item-level analysis of results. For example:

> Teachers are expected to get into their professional learning communities, and have conversations about the results. And there's been a protocol developed that they should use to make sure that they—step one—just make factual statements about the results … eventually they'll get into interpretation... They may say things like, 'Well, it would make sense that 19% of these questions were answered correctly because we didn't spend much time on them,' or whatever. And then, the next step would be for them to identify a focus. "Okay, we're going to focus on [the] [number] [sense] standard..." And then they will have a conversation about how they're going to adjust their instruction to support focusing on that particular content, with even a timeline ***(Sinclair SD, Assessment Director)***.

In contrast with administrators' reports, teachers in only two districts mentioned participation in team meetings related to the assessments. In Madison SD, teachers described data teams focused on reteaching items on which students scored poorly, with only one teacher giving an example of collaborating around diagnostic multiple choice item responses and in turn creating classroom activities. In Burlington SD, team meetings consisted of "structured teacher planning time" in which content-area teams analyzed assessment results and engaged in collaborative planning to ensure coverage of the standards, especially those that would appear on the state test. Thus, despite the fact that all but one district had been using its interim assessment for at least three years, only Madison SD and Burlington SD showed evidence of ongoing teacher collaboration around results.

Build Teachers' Content Knowledge. In three districts (Burlington, Madison, and Sinclair), district-level administrators wanted PD efforts to support teachers' planning specific to knowledge of the content standards. In Burlington SD, "structured teacher planning time" was provided, as described above. Administrators in the other two districts intended to increase teachers' content knowledge by emphasizing the "big ideas" for each concept, suggesting a lesson structure, or tailoring PD to specific "units of study." As one PD Director noted:

> Our primary goal is increasing teacher content knowledge and then pedagogy. Our theory of action in the district… is that it's the teacher's expertise that has the greatest impact on increased student achievement, so …what we're doing with our pacing guides is now providing professional development that only covers a small section of it and then we schedule the next piece to cover the next section... ***(Madison SD)***.

Madison SD's pacing guide provided a structure that informed content-focused PD throughout the year, which promoted the district's intentions of both ensuring consistent coverage and augmenting teachers' expertise in the standards.

Set Goals for Student Growth and Target Disparities in Achievement. Some district-level administrators described PD focused on setting goals for the academic growth of all students, sometimes in concert with individual student feedback. Washington SD focused on "learning targets," by which students would be aware of lesson objectives and ways to demonstrate learning. The Pittsfield and Taylor districts aimed for teachers to understand what would be "reasonable growth targets" for students in terms of improving state test scores.

Principals focused on results disaggregated by race, gender, and English learner status in order to address disparities in achievement. One principal recounted her school's implementation of what she called "culturally responsive pedagogy" (e.g., Ladson-Billings, 1995; Santamaria, 2009). In practice, however, students were assigned membership in gendered "Black, Hispanic, and White Leadership Groups" and encouraged to improve that group's test results on specific standards. Other schools offered PD on scaffolding math instruction for English learners.

Provide Specific Instructional Strategies. In Burlington SD, teachers were encouraged to use an assessment report that grouped "mastered" and "non-mastered" students within each standard in order to "differentiate" instruction by offering differently leveled problems. However, most teachers who discussed PD on informing instruction described receiving guidance on specific instructional tactics that were not explicitly connected to the use of interim assessment results. Teachers in Madison SD, for example, were trained to use a "number talk" strategy (separate from the interim assessment) in which teachers interviewed students on a series of progressively more challenging problems related to numeracy and computation.

Provide Technical Training. The few district-level administrators that reported providing PD on technical aspects of the assessment system noted the need for this in the early stages of implementation, with later efforts progressing to using the results to inform instruction.

Though fewer than a third of administrators discussed this type of PD, the great majority of teachers across districts reported technical training as the primary form of staff development that they had received. They did not describe a transition to more substantive PD, explaining instead that training was offered in the first years of the assessment, but then tapered off. The following is a typical teacher response in this category:

> The first year it was implemented we were given trainings [on] [the] [software]. So, thankfully I was here, and it was a 45-minute training where they went in and showed us what the website offered and how to access all of your students' information... However, if you weren't here that first year it was implemented, there hasn't been another training. … As far as professional development, that's all that's been offered to me ***(Madison SD, Elizabeth)***.

Teachers in Sinclair SD noted that for the previous district assessment which had consisted of all constructed response items, teachers had engaged in PD in which they reviewed various student approaches to problems and agreed upon grading techniques. After the district switched to mostly multiple choice with a few "enhanced multiple choice" items - due to the fact that some students had difficulty in explaining their answers and this format did not reflect what was required on the state test - PD on the assessment was no longer viewed to be necessary.

Research Question 3

What actual uses of interim assessment results did stakeholders report (i.e., changes in instruction, improved course pacing, and/or enhanced collaboration)? Were the reported uses coherent (i.e., aligned with the original purposes and expectations and with uses reported by others)?

We asked all 24 district-level administrators and 14 principals to describe how assessment results were being used at the district, school, and classroom levels, and probed for specific examples.[8] To enable comparisons with the findings on district intent, we present respondents' descriptions in terms of evaluative, instructional, and predictive practices, and note variation in perspectives held by top-level and frontline administrators. We again indicate salient differences in frontline administrators' responses for districts with only multiple-choice exams versus those including some constructed response items.

[8] Teachers' descriptions of their assessment practices are reported in detail in Shepard et al. (2011).

Overview. Multiple assessment purposes and practices were reported in the administrator interviews and treated as compatible. Overall, informing instruction seemed to be at the core of assessment use, with the state standards determining the data that would drive the system. District-level administrators and principals generally gave similar reports of instructional practices related to the assessments, such as broadly informing instruction, providing feedback, or grouping. Principals additionally described a use of results for student placement into leveled classes. While most district-level administrators also cited evaluative practices, such as monitoring achievement for accountability purposes or informing needs for PD or resources, only a few principals made similar claims. Predictive practices were rarely mentioned, and typically coupled with other instructional practices. Table 4 summarizes reports of practices.

Table 4

Administrators' and Principals' Understandings of Interim Assessment Practices

Practices	"Top-level" administrators *(N=5)*	"Frontline" administrators		Principals (*N*=14)
		MC[a] only (*N*=11)	MC + CR[b] (*N*=8)	
Instructional	100%	73%	88%	80%
[c]C:"Inform instruction"	100%	64%	38%	86%
C: Goal setting/feedback	20%	9%	13%	21%
C: Differentiation/grouping	40%	0%	13%	43%
C: Placement	0%	0%	13%	43%
[d]S/D: Assist with data analysis and/or data-driven dialogue	20%	27%	63%	29%
Evaluative	80%	91%	88%	64%
S/D: Educator evaluation	20%	9%	13%	14%
[e]D: Monitor accountability	60%	55%	38%	57%
D: Inform school level needs for PD and/or resources	60%	82%	88%	14%
Predictive	20%	9%	25%	29%
S/D: Predict state test	20%	9%	25%	29%

Note. [a]MC = multiple choice; [b]CR = constructed response; [c]C=classroom-level uses; [d]S/D=school/district-level uses; [e]D= district level uses.

Instructional Practices. Administrators in districts with constructed-response items emphasized efforts to initiate teacher collaboration through data-driven dialogue, while those in districts with only multiple choice items more often broadly stated that data was being used to "inform instruction." Only a few administrators named specific practices, such as providing students with feedback on their scores and grouping students by performance on the standards.

The ability to potentially identify students needing remediation was commonly understood by administrators as a benefit of using a benchmark assessment system. Four district- and school-level administrators in different districts (Burlington, Sinclair, Taylor, and Washington) specifically referenced using assessment results as part of the *Response to Intervention* (RtI) strategy,[9] with several others using similar language of targeting students needing intervention:

> You've heard of the *Response to Intervention* pyramid? Okay, we're working right now mostly with that first bottom tier, building a common strong, standards-based, relevant, rigorous core for everybody... and then... we're going to be talking about... Tier 2, what do we push in? How do we identify which kids are struggling with what? How do we group them or support them or scaffold for them? So, we've got this plan that brings together curriculum, instruction, assessment, and we're trying to build it this year using that frame of the *Response to Intervention* pyramid ***(Sinclair SD, Curriculum Specialist)***.

The aim of RtI is to make instructional modifications based on information about student learning in different contexts in order to avoid over-identification of learning disabilities. These administrators implied that interim assessment results would be one gauge of whether instructional strategies had succeeded with particular students within the overall RtI strategy.

Some district-level administrators expressed concern that data-driven instruction was not occurring as much as it should be, while others pointed out that its implementation varied widely among teachers. For example:

> ...we haven't gotten to the point where the teachers are using the data from these assessments to really inform their instruction and to work together... That's where we want to go and we've been at it for... six years and we haven't gotten there ***(Burlington SD, Superintendent).***

> It varies. I can't say that every teacher values the benchmark test... Our teachers are just like any school district. They are everywhere in terms of their technology level and willingness to use data. But we see improvement over the years ***(Adlington SD, Curriculum Specialist)***.

[9] For more on RtI, see Fletcher and Vaughn (2009), Fuchs and Fuchs (2006), and Samuels (2011).

Some principals expressed similar concerns, but were typically more familiar with specific instructional practices. A few described teachers giving feedback to students based on overall scores and proficiency in the standards:

> We've really pushed our kids this year to know what their NWEA scores are. And if you walk around the building we have charts on the wall [and] [in] [every] [classroom] that will say, "Where do you score?"... And it's got like an NWEA reading and NWEA math. And it will say, "Sixth grade, you need to score here. Seventh grade. And eighth grade." So the kids see that in front of themselves all the time ***(Washington SD, Principal B)***.

Almost half of principals cited within-class differentiation (largely through grouping) and student placement in leveled classes, with a great deal of overlap occurring for these two practices. Grouping strategies were typically informed by performance on standards or items. (In Madison SD, this was facilitated by score reports that listed groups of "mastered" and "non-mastered" students for each standard). Teachers either led small group instruction, encouraged peer tutoring, or asked students to work collaboratively:

> So in the best case scenarios teachers are using the *DesCartes Continuum*[10] to form groups. To look at acceleration for students who perhaps have mastered something, to look at remediation for kids who need one strand to be bolstered ***(Washington SD, Principal A).***

> [The] [teachers] do use the results. It helps them with instruction, in forming groups, okay? If they see that a child is having difficulty with a certain concept, they can put the children together and work with that child on that concept. And then they do like a little post-test and then pull that child out of that group and maybe pull in some more. So the groups have to be flexible, okay, but the tests do inform the instruction of the teachers ***(Adlington SD, Principal A).***

A few principals noted that some teachers examined constructed response items for a deeper understanding of students' knowledge:

> I believe that teachers do look at [the] [constructed] [response] [items] because they want to see what the thinking that the kids had to lead them to the answer. But if the teachers don't want to do that, they don't have to ***(Sinclair SD, Principal A).***

Student placement in leveled classes, however, was typically based on overall scores, and was especially tied to the first and last test administrations of the year:

> First and foremost they start out with a benchmark test to figure out where each student is and then we can place students accordingly. If they... are in need of higher level math, then

[10] As part of the NWEA MAP® assessment, the *DesCartes Continuum* lists skills and concepts to focus on by standard and RIT score range.

> we can place them via that first entry level test. If they still haven't mastered the basics, then we can place them appropriately, as well ***(Pittsfield SD, Principal A).***

Because of the focus on proficiency in the standards, feedback and grouping practices could be seen as compatible with district-level administrators' interests in aligning curriculum and informing instruction. On the face of it, however, the widely cited use for student placement does not appear to be consistent with the original purposes of assessment adoption.

Evaluative Practices. Monitoring achievement for accountability was a frequently cited evaluative practice. Principals were typically briefed on school-level comparisons at district meetings or through distributed reports, which also included attention to differences in results among English learners and other sub-groups. Several respondents noted that results were also being used for educator evaluation:

> I'm evaluated as a principal based on my data... We use this as part of the evaluation process. If we have a teacher that's showing 32% targeted gains, there's a good chance that that teacher – for second- and third-year teachers, they'll be non-renewed ***(Taylor SD, Principal A).***

Finally, district-level respondents claimed that assessment results were used to inform school-level needs for PD or other resources, but few principals noted this practice.

Predictive Practices. Several respondents noted *predictive* uses of interim assessment results at the district and school levels to anticipate and improve state test performance:

> We, of course first, analyze our [state] [test], look at the student needs. And then we look at the alignment piece from the [benchmark] [assessment] to the [state] [standards].... And we select focus standards to teach... to mastery. And then, of course, we analyze it... So we have these guiding questions that we use to re-identify the next group of standards that we're going to be teaching from the benchmark assessments, but that they're also aligned and they're going to be tested on [the] [state] [test] ***(Burlington SD, Principal B)***.

Although evaluative and instructional practices were much more frequently cited, improved performance on the state test was more salient in actual practice than it had been in explanations of purpose.

District-by-District Patterns of Coherence

Given the importance placed on clear and shared understandings of the purposes of an assessment to ensure test validity (e.g., Perie et al., 2009), we sought to understand district coherence by analyzing each district's respondents' perspectives as a set. We found fairly consistent understandings of assessment purposes and practices among district-level administrators. The extent to which these understandings were shared by principals and teachers

in each district, however, varied greatly. When asked "how the assessment was selected and implemented," almost all principals claimed that the district simply chose an assessment and mandated its use. Teachers were likewise uncertain of the district intent, such that both principals and teachers appeared to glean their understandings from assessment-related practices at the district, school, and classroom levels. In addition to comparing various respondents' views of district intent, we examined the coherence of intent with the actual practices they described. Here we discuss these forms of coherence by presenting a picture of each district.

In Burlington SD, district-level administrators emphasized the need to "shock the system" into establishing a common curriculum that aligned with the state standards and state test. The need to standardize the curriculum reflected a primary aim on the part of administrators in this district when they talked about informing instruction. For this reason, many regretted the lack of open-ended items on the tests. Only the Assessment Director described practices of training administrators as instructional leaders and using test results for student placement. However, all administrators referred to the need for teachers to develop content knowledge. Burlington therefore offered content-specific PD, academic coaches, and "structured teacher planning time" (STPT). During STPT, teachers jointly planned lessons specific to content strands. Although the Superintendent suggested that most teachers were not yet successfully using the results to inform instruction, both principals and most teachers described STPT as valuable for their own instructional improvement. At the same time, no teachers mentioned instructional purposes as a district intent and instead varied in their claims of evaluative and predictive aims. The district also engaged in evaluative practices of monitoring accountability, comparing schools, and evaluating principals.

District-level administrators in Madison SD saw the interim assessment as a means to promote consistency in teachers' adherence to the already-in-place district curriculum and pacing guide. Because they hoped that the assessment would be used to inform instruction, the district included diagnostic multiple choice and constructed response items on the tests, and scheduled time for data team meetings. Principals and teachers likewise discussed the use of data teams in schools. However, principals placed more emphasis on the district desire for accountability and state test prediction - which was only secondarily mentioned by district-level administrators - and teachers varied greatly in understanding the primary purpose of the interim assessment as instructional, evaluative, or predictive. Only one teacher reported analysis of the diagnostic multiple choice answers, while others claimed the scoring of the constructed response items was most useful for informing their instruction. Most teachers reviewed the assessment, item by item, with the whole class and engaged in small-group instruction on missed items. While the district did not explicitly connect teacher training on using "numeracy talks" to interim assessment

results, one principal claimed that teachers' regular use of this strategy in the classroom had a strong impact on students' mastery of numeracy and other standards.

Adlington SD also had a curriculum in place at the time of benchmark assessment adoption, but did not have a pacing guide. District goals were then to raise an awareness of the standards among teachers, and ensure coverage of the curriculum and appropriate pacing. In this way, district-level administrators viewed the benchmark as a summative gauge of whether students had mastered the content. Tests included one constructed response item to emphasize that students should be problem-solvers, and there was some interest in reteaching standards with low scores. To this end, district-level administrators identified low-performing standards across the district, and provided PD on instructional strategies specific to that content. However, both district-level administrators and principals emphasized the purpose of the assessment as ensuring that teachers were covering the standards. While the district shared strategies such as teacher collaboration at principal meetings, instructional practices were considered a site-based decision, and varied by school. Both principals indicated that the district simply mandated the assessment and were not familiar with teacher participation in related PD. Both also noted that some teachers were reviewing results in grade level meetings and using results to group students within classes, while one claimed that results were being used for student placement.[11]

Pittsfield SD did not have a district-wide curriculum but indicated that for mathematics, "the textbook was the curriculum." The district chose to use the quarterly assessments provided by the textbook, but contracted with an online data collection and reporting company because it would "provide the leverage" to establish a common curriculum. District administrators hoped that once teachers were on the same page with regard to the curriculum and standards, they would collaborate in team meetings on using test results to inform their instruction. To accomplish this, the district offered an "inquiry protocol" three years prior to our study; lead teacher training for data team processes; principal training for instructional leadership; and dedicated staff development time for data team meetings every other week. However, the actualization of these practices was considered a site-based decision, and the extent to which data teams and instructional leadership characterized school practices varied. Neither principal described a role of instructional leadership. One principal criticized the district's support for implementing the system, including a lack of PD. In that school, test results were being used for student placement, to inform the purchase of intervention programs, and in grade level team meetings to target areas for review. For example, test results were used to determine a school-wide, pre-state test focus on specific basic skills (such as memorizing multiplication tables). The

[11] We did not include questions on district intent with teachers in this district, so cannot compare their perspectives here.

other principal likewise mentioned the use of results for student placement, but noted district involvement in implementation via administrator classroom walk-throughs and PD. He regretted that the instrument did not yet provide enough information to inform instruction. While he encouraged teachers to analyze test data, he described his job as providing the resources for teachers to learn from one another (such as providing substitute teachers so that teachers could observe each other). Most teachers cited a primary district aim of monitoring curriculum coverage and pacing, with others naming instructional and predictive goals. None of the teachers reported the use of assessment results or a protocol in team meetings.

Taylor SD had previously used the paper-and-pencil "Levels" format of the NWEA MAP® with students who were not proficient on the state test. Conversion to the computer-adaptive version was being piloted with the intention of expanding the use of the assessment to all students. Administrators noted some interest in monitoring aggregated data and predicting state test scores, but emphasized an instructional intent. Because curriculum was previously a site-based decision, the district aimed to define common standards, curriculum, and learning targets through the assessment. Administrators expressed hope that the DesCartes tool would inform instruction, but recognized the need to provide PD because "it's not always clear to teachers about where I need to go next with a particular group of kids." One noted concern with the multiple-choice only format because of the importance of "understanding kids' thinking and why they arrived at the answers that they did and trying to make sure that we're going beyond repeating algorithms to really application and transfer." However, the two principals whom we interviewed indicated different purposes and practices for the assessment. One principal emphasized a focus on the achievement gap to such an extent that all student groupings were informed by disaggregated results by race and gender. This school also used the test for student placement and teacher evaluation, and the principal indicated that test results were part of her principal evaluation. At the same time, she noted that the school was moving toward more frequent formative assessment and collaborative learning. The other principal hoped to use the test "formatively," but had not yet considered teacher collaboration or structured use of the results. She was largely unaware of both the district's and her own teachers' practices related to the assessment, and was doubtful that many teachers knew how to use data. Most teachers in this district thought the assessment was primarily intended to predict the state test, with a few others citing instructional and evaluative aims.

Washington SD had also been using the paper-and-pencil version of the NWEA MAP® "Levels" test for several years, and had recently upgraded to the computer-adaptive version. The goals of the assessment remained primarily focused on accountability and state test prediction. However, with the DesCartes instructional tool, administrators also expressed hope that teachers

would use test results to inform instruction. The district trained test proctors (often paraprofessionals) for administration and data facilitators (often lead teachers) for data analysis. Principals were expected to base school improvement plans on student performance, and determine the appropriate needs for PD and intervention programs. At the same time, Washington SD had just adopted a new textbook, and was in the process of redefining the district standards, creating a pacing guide, and seeking a new benchmark assessment system that would better align with the state standards and state test.[12] Principals did not discuss this transition, but both corroborated the emphasis on state test prediction, the use of results in school improvement plans, the involvement of test proctors and data facilitators, and teachers' use of the DesCartes tool for student grouping. Both also used results for student placement and displayed proficiency charts throughout the school for students to receive feedback on scores. Interestingly, all teachers understood the district's aims for the assessment to be gauging student mastery of content and informing instruction, and none mentioned district goals related to accountability or state test prediction. Only two of six teachers echoed the use of grouping practices that their principals had described.

Finally, Sinclair - a very large district - emphasized site-based decision-making to such an extent that district-initiated practices were enacted to widely varying degrees across schools. Sinclair had implemented a common curriculum several years prior through district-created modules that included lesson plans and end-of-unit assessments. The district then created quarterly benchmark assessments, but emphasized to teachers that the tests were to be used to inform instruction, and would not have to be reported to the district. In order to promote a "common conversation," however, Sinclair later modified the benchmark test to more closely align with the state test, and required that results for the multiple choice items be sent to the district. The "enhanced multiple choice" items remained optional for teachers to analyze, such that teachers' attention to these items varied greatly. While district-level administrators and principals described professional learning communities as the primary medium for teacher use of assessment results, no teachers mentioned this practice. Instead, most teachers saw district aims as related to accountability.

Summary. While recognizing that this analysis is limited by the small sample size of the teacher and administrator groups within each district, we note general patterns. First, there was general coherence among district-level administrators within each district with respect to assessment purposes. In light of this, principals' and teachers' high degree of uncertainty of district goals is surprising. This juxtaposition strongly suggests that the purposes of the

[12] In fact, Washington SD began using Scantron in the 2009-2010 school year (School Board Meeting Minutes, http://www.boarddocs.com/co/acsd50/Board.nsf/Public).

assessments intended by those involved in their adoption at the district level were not communicated on a clear and consistent basis at all levels within the district, resulting in a lack of shared understandings.

We also found varying degrees of coherence of intended purposes with assessment practices. While instructional aims were emphasized in most districts, in only three districts did administrators purposefully include constructed response items to this end. In three of the other four districts, district leaders consistently regretted the entirely multiple choice format of the tests. Likewise, administrators suggested that teachers were not skilled in using data to inform instruction, but the only thorough-going instances of staff development to this end were the structured team meetings in Burlington and Madison. While some districts used a trainer-of-trainer model, it did not seem to be effective in that few teachers reported learning how to use assessment results from academic coaches. Districts often cited a site-based decision making culture that left it to principals to provide the support teachers needed to effectively use assessment results; this approach resulted in varying levels of coherence of school-level practices with district-level intent.

Discussion

The findings of this study are generally consistent with the existing research literature on interim assessments and data-based decision making. Top-level administrators, frontline administrators, principals, and teachers reported different understandings for the purposes and expectations of interim or benchmark assessment. Implementation issues, such as lack of capacity, impacted use of the assessment results. Finally, the assessments were sometimes used in ways that had not been anticipated in the adoption process. We summarize each of these findings in turn, and then make some overarching observations and recommendations.[13]

First, the *purposes and expectations* of the interim assessments varied among different stakeholders. For example, 75% of teachers did not know or were unsure of their district's intentions with respect to the assessments, suggesting weak communication to the teachers from district or school administrators. Coherence among district-level administrators themselves was better: Though various administrators reported a wide variety of purposes within each district, they largely agreed on either instructional or evaluative categories of purpose. The breakdown in communication of the purposes of the assessment from the central office to the schools was apparently accompanied by a breakdown in subsequent implementation, including professional

[13] Because our sample was unrepresentative, our results should not be directly generalized to all district contexts. However, our findings do add to the literature base in a way that should inform research and policymaking (see Eisenhart (2009) for more on generalization in qualitative research).

development. The lack of shared understandings of the purposes of the assessment also led to somewhat divergent uses of assessment information.

Second, though administrators acknowledged the importance of professional development to train teachers in the use of assessment results, districts did not follow through with building capacity in this way. Here, several disconnects seemed to occur. In most cases, administrators described PD efforts focused on encouraging data-driven dialogue in schools, but few teachers reported participating in this type of training. At the same time, administrators overwhelmingly cited a lack of teacher knowledge in effectively using assessment results as a major weakness of the system, as well as wide variation in the extent to which individual teachers used the assessment results. However, they did not appear to connect these issues to a lack of PD. Principals reported additional staff development that directly guided planning, goal setting, and instructional strategies, which was corroborated by some teachers. Yet, most teachers claimed to have received only technical training or not much PD at all. The Burlington Superintendent, who cited a lack of accountability with respect to PD, gave one possible explanation for these breakdowns: "So, who monitors and keeps teachers accountable and principals accountable? There was no one. So imagine all this professional development and there was no accountability."

Third, adoption processes did not anticipate the variety of ways in which assessment results would be used. For example, the most common assessment format—multiple-choice—was consistent with evaluative aims such as aggregation of achievement data, but limited instructional decision-making to inferences about students' performance based on standard- and item-level information. Some districts' efforts to include constructed response items with the intention of informing instruction were hindered by both a short turnaround time that limited teachers' opportunity to review and grade these items and an emphasis on the aggregated results from the multiple choice items. Not surprisingly, then, the instructional goals and uses of the assessments in most districts reflected only a vague expectation that teachers would respond instructionally to assessment results in order to promote student mastery of the content standards.

Importantly, an unanticipated assessment use was the placement of students in leveled classes based on overall scores.[14] As Oakes (1990), Braddock et al. (1993), Burris, Heubert, and Levin (2004), and others document, tracking can have important and negative consequences for students' continued opportunities to learn mathematics. While Stiggins (2006) notes that American schools have abandoned the historical system of sorting students by perceived ability

[14] Ash (2008) likewise notes some schools' use of results from the NWEA MAP® computer adaptive assessment to group students within classrooms and into leveled classes.

in favor of a view that all students can learn, it is possible that assessment data used for placement in this way could similarly sort students. The National Research Council (1999) cautions against important decisions being made from a single assessment, and Li et al. (2010) note that ability grouping based on one assessment should be considered "at least moderate stakes" (p. 169) with a corresponding need for safeguards. The use of benchmark test scores to identify students needing tiered levels of remediation as part of the *Response to Intervention* strategy has likewise concerned experts in the field of special education (Fletcher & Vaughn, 2009; Fuchs & Fuchs, 2006).[15]

Researchers have established a collective vision of interim assessment that could effectively inform instruction as well as serve evaluative purposes, but this would require substantial improvement in the nature of the instruments. According to Perie et al. (2009): item formats would be varied; insights into learning would be possible through open-ended items; instructional strategies would be proposed in response to substantive insights; test content would be aligned or "linked" to the curriculum; exams would not interrupt the learning of curriculum but rather be integrated; and related PD would be provided (p. 10). Herman and Baker (2005) and Shepard (2005) echo the need to include constructed response items, and to ensure that assessments reflect the full range of the content of interest through "content mapping" and attention to "big ideas" rather than mimicking current state tests. Though the assessments implemented in the districts we studied generally fell far short of this ideal, those that included constructed-response items came closer than those that relied on multiple-choice items alone. Indeed, administrators in districts with constructed-response items claimed that these item types were a strength of the system, while administrators in districts with only multiple choice items regretted that they could not obtain more specific and diagnostic information.

With respect to this vision of interim assessment, Madison SD was exemplary in our sample. Open-ended and diagnostic multiple-choice items provided opportunities to obtain insight into students' understanding as well as aggregate achievement data to track progress toward meeting instructional goals. Though teachers did not always make use of test results instructionally, at least some engaged in data-driven dialogue facilitated by district PD. Test content was linked to the classroom text, district curriculum, and state standards; thus, the test appeared to connect learning targets at various levels. Clearer communication by district administrators about the intended purposes of the assessment and the importance they placed on learning about students' understanding through the constructed-response items, along with

[15] For example, NWEA advertises the computer-adaptive test used by two districts in our study as useful for RtI, state test prediction, curriculum alignment with state standards, and formative assessment (see http://www.nwea.org/help-all-kids-learn). Other benchmark assessments make similar claims.

providing time and increased guidance to teachers for this purpose, could render this interim assessment even more effective instructionally.

In order to for interim assessments to be used effectively, we emphasize that true district coherence revolves around a *well-developed, shared, research-based theory of action*. As Fullan (2011) notes, it is essential that the "right drivers" lead "whole system reform" (p. 3). Fullan argues that initiatives that prioritize "accountability, individual teacher and leadership quality, technology, and fragmented strategies" (p. 5) serve to decrease motivation and hinder success. On the other hand (p. 6):

> The right drivers – capacity building, group work, instruction, and systemic solutions – are effective because they work directly on changing the culture of school systems (values, norms, skills, practices, relationships); by contrast the wrong drivers alter structure, procedures and other formal attributes of the system without reaching the internal substance of reform – and that is why they fail.

In this way, top-down pressures that drove the adoption of interim assessment systems in most of the districts in our study seemed to lay an accountability-based reform on top of existing practices, which undermined administrators' expressed interests in improving instruction and fostering collaboration.

The current national policy climate, however, offers an important opportunity to clarify the purposes of various assessment systems and promote coherence in their use. The federal government has funded two consortia of states - the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC) - to design assessment systems that are aligned with the common core state standards and are intended to serve purposes of accountability *and* instructional improvement. These goals mirror the tensions we found in districts' aims for interim assessments. Contrary to the assessments in our study in which representations of learning goals were limited, both consortia have emphasized the importance of interim assessments that include multiple item types offering substantive information on students' understanding, along with PD that promotes teachers' collaborative use of results to inform instruction.[16]

Key Recommendations

Drawing from our findings and other research, we offer six recommendations for a successful interim and benchmark assessment system. Though having these conditions in place does not guarantee the success of the system, we believe they are essential as a minimum.

[16] SBAC also states that interim assessments should be optional for districts.

1. Assessment systems should be based on a well-developed theory of action, including clearly defined purposes, expectations, and uses. Great caution is urged in adding any new assessment program on top of existing programs.
2. A written plan developed by and reflecting perspectives from different stakeholders should be created to identify program goals, understandings of purposes, and assessment coherence; the plan should address appropriate use of assessment results.
3. Assessment selection should include multiple stakeholders; moreover, assessments themselves should include a combination of high quality selected-response and constructed-response items aligned to standards, curriculum, and textbooks.
4. Sufficient professional development must be provided in order for teachers to engage in data-driven dialogue, adjust their instruction, and evaluate its impact; this PD should be explicitly linked to other PD efforts.
5. Systematic communication should be planned for and carried out among participants at all levels, regarding the purposes and appropriate uses of the assessment and the progress of ongoing professional development. This communication must be regular and sustained over time for continued coherence of the assessment system.
6. Internal, or ideally external, ongoing evaluation of program processes and effects is necessary to ensure implementation fidelity and effective use of the assessment system to support instructional improvement and student learning.

While our recommendations may be helpful to school districts and schools as they implement interim or benchmark assessment systems, the PARCC and Smarter Balanced assessment consortia can be instrumental in designing such programs, and more generally, in helping school districts build the shared understandings of assessment purposes that are necessary for valid and coherent assessment use.

References

Ash, K. (2008, November 19). Adjusting to test takers. *Education Week.* Retrieved from http://www.nwea.org/about-nwea/news-and-events/adjusting-test-takers

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice.* Berkshire, UK: Open University Press.

Black, P. J., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, & 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.

Braddock II, J., Hawley, W., Hunt, T., Oakes, J., Slavin, R., & Wheelock, A. (1993). *Realizing our nation's diversity as an opportunity: Alternatives to sorting America's children.* (Final Report to the Lilly Endowment). Washington, DC: Common Destiny Alliance.

Burris, CC. & Heubert, J.P., & Levin, H.M. (2004). Math acceleration for all. *Educational Leadership, 61* (5) 68-71.

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

Coburn, C.E. & Talbert, J.E. (2006). Conceptions of evidence use in school districts: Mapping the terrain. *American Journal of Education, 112,* 469-495.

Copland, M.A., Knapp, M.S., & Swinnerton, J.A. (2009). Principal leadership, data, and school improvement. In T.J. Kowalski & T.J. Lasley II (Eds.), *Handbook of data-based decision making in education* (pp. 153-172). New York, NY: Routledge.

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

DuFour, R. & Eaker, R. (1998). *Professional learning communities at work: Best practices for enhancing student achievement.* Bloomington, IN: National Education Service.

Eisenhart, M. (2009). Generalization from qualitative inquiry. In K. Ercikan & W.M. Roth (Eds.), *Generalizing from educational research: Beyond quantitative and qualitative polarization* (pp. 51-66). New York, NY: Routledge.

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.

Erickson, F. (1986). Qualitative methods in research on teaching. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 119-161). New York, NY: MacMillan.

Fletcher, J.M. & Vaughn, S. (2009). Response to Intervention: Preventing and remediating academic difficulties. *Child Development Perspectives, 3*(1), 30-37.

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, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Fuchs, D. & Fuchs, L.S. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly, 41*(1), 93-99.

Fullan, M. (2011, May). *Choosing the wrong drivers for whole system reform*. (Seminar Series Paper No. 204). Melbourne, Australia: Centre for Strategic Education.

Gewertz, C. (2010, March 3). Testing experts lay out vision for future assessments. *Education Week.* Retrieved from http://www.edweek.org/ew/articles/2010/02/23/23assessment.h29.html

Goren, P. (2010). Interim assessments as a strategy for improvement: Easier said than done. *Peabody Journal of Education*, *85*(2), 125-129.

Halverson, R. (2010). School formative feedback systems. *Peabody Journal of Education*, *85*(2), 130–146.

Hattie, J.A.C. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement.* Abingdon, UK: Routledge.

Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity?* Paper prepared for the Council of Chief State School Officers. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J. L. & Baker, E. L. (2005). Making benchmark testing work. *Assessment to promote learning*, *63*, 48–54.

Herman, J.L. & Gribbons, B. (2001). *Lessons learned in using data to support school inquiry and continuous improvement: Final report to the Stuart Foundation* (CSE Technical Report 535). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J.L., Osmundson, E., & Dietel, R. (2010). *Benchmark assessment for improved learning* (AACC Policy Brief). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Honig, M.I. (2003). Building policy from practice: District central office administrators' roles and capacity for implementing collaborative education policy. *Educational Administration Quarterly, 39*(3), 292-338.

Kerr, K.A., Marsh, J.A., Ikemoto, G.S., Darilek, H., & Barney, H. (2006). Strategies to promote date use for instructional improvement: Actions, outcomes, and lessons from three urban districts. *American Journal of Education, 112*(4), 496–520.

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Education Research Journal, 32*(3), 465-491.

Li, Y., Marion, S., Perie, M., & Gong, B. (2010). An approach for evaluating the technical quality of interim assessments. *Peabody Journal of Education*, *85*(2), 163–185.

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.

Marsh, C.J. (2007). A critical analysis of the use of formative assessment in schools. *Educational Research, Policy, and Practice, 6*, 25-29.

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

McLaughlin, M. W., & Mitra, D. (2003). *The cycle of inquiry as the engine of school reform: Lessons from the Bay Area School Reform Collaborative*. Stanford, CA: Center for Research on the Context of Teaching.

Miles, M.B. & Huberman, A.M. (1994). Qualitative data analysis: An expanded sourcebook, (2nd ed.). Thousand Oaks, CA: Sage.

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.

National Research Council (1999). *Testing, teaching, and learning: A guide for states and school districts.* Washington, DC: National Academy Press.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Oakes, J. (1990). *Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science.* Santa Monica, CA: The RAND Corporation.

Partnership for Assessment of College and Career Readiness (PARCC) (2011). Accessed Retrieved from http://www.parcconline.org/.

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.

Picciano, A.G. (2009). Developing and nurturing resources for effective data-driven decision making. In T.J. Kowalski & T.J. Lasley II (Eds.), *Handbook of data-based decision making in education* (pp. 123-135). New York, NY: Routledge.

Popham, W.J. (2006). A tale of two test types. *Principal*, March/April, 12-16.

Porter, A. (1989). A curriculum out of balance: The case of elementary school mathematics. *Educational Researcher, 18*(5), 9-15.

Ryan, K.E. & Shepard, L.A. (Eds.). (2008). *The future of test-based educational accountability*. New York, NY: Routledge.

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

Samuels, C.A. (2011, February 28). An instructional approach expands its reach. *Education Week.* Retrieved from http://www.edweek.org/ew/articles/2011/03/02/22rti-overview.h30.html?tkn=YPNFtRuk%2F4QwJ2dSrpcmU%2F4Uk9qzp%2FdwZYno&cmp=clp-edweek?Intc=RTI11EWH.

Santamaria, L.J. (2009). Culturally responsive differentiated instruction: Narrowing gaps between best pedagogical practices benefiting all learners. *Teachers College Record, 111*(1), 214-247.

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

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

Shepard, L.A. (2010). What the marketplace has brought us: Item-by-item teaching with little instructional insight. *Peabody Journal of Education*, *85*(2), 246-257.

Shepard, L., Davidson, K., & Bowman, R. (2011). *How middle school mathematics teachers use interim and benchmark assessment data*. (CSE Technical Report). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1–22.

SMARTER Balanced Assessment Consortium (SBAC) (2011). Retrieved from http://www.k12.wa.us/SMARTER/.

Spillane, J. P., Halverson, R. R., & Diamond, J. B. (2001). Investigating school leadership practice: A distributed perspective. *Educational Researcher*, *30*(3), 23–28.

Stiggins, R. (2006, May). *Balanced assessment systems: Redefining excellence in assessment*. Portland, OR: Educational Testing Service.

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

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

Wayman, J.C., Midgley, S., & Stringfield, S. (2005). *Collaborative teams to support data-based decision making and instructional improvement.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Wayman, J.C. & Stringfield, S. (2006). Technology supported involvement of entire faculties in examination of student data for instructional improvement. *American Journal of Education, 112,* 549-571.

Wenger, E. (1998). *Communities of practice: Learning, meaning and identity*. New York, NY: Cambridge University Press.

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

Wiliam, D. (2007). Keeping learning on track: Classroom assessment and the regulation of learning. In F.K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1053-1098). National Council of Teachers of Mathematics: Information Age Publishing.

Young, V.M. (2008). Supporting teachers' use of data: The role of organization and policy. In E.B. Mandinach & M. Honey (Eds.), *Data-driven school improvement: Linking data and learning* (pp. 87-106). New York, NY: Teachers College Press.

Young, V. M., & Kim, D. H. (2010). Using assessments for instructional improvement: A literature review. *Education Policy Analysis Archives, 18* (19). Retrieved from http://epaa.asu.edu/ojs/article/view/809

Appendix A:
Administrator Interview Protocol

Notes to Interviewers:

Who is to be interviewed

Depending on the size and organization of the district, the following people will be interviewed:
• Math curriculum coordinator, or other person responsible for middle school math curriculum
• Assessment coordinator, or other person responsible for assessment
• Professional development coordinator, or other person responsible for professional development related to middle school math and to assessment
• Superintendent and/or deputy superintendent

Notes regarding the content and format of the template

A version of these questions will be posed to all district-level personnel to be interviewed.

Topic headings, not to be read, are in small caps. Probes for some of the questions, to be used as necessary based on the initial responses, are italicized.

Questions surrounded by asterisks pertain to factual data or sample materials, which will be collected prior to the interviews if possible. Where this is not possible, the questions will be asked of the various respondents only until the information is obtained, then stricken from the protocol for the rest of the district's respondents.

Questions to be asked of only a subset of interviewees from a district are indicated by a shaded statement. Some topics contain "bail-out" questions, also shaded, to provide the respondent with the opportunity to indicate that he or she is unfamiliar with the topic, in which case the rest of the questions on that topic are omitted. The bail-out question can be skipped if it is clear to the interviewer that the respondent is familiar with the topic.

Other than the types of questions mentioned above, each question should be asked of every respondent for the district.

Preparation and Introduction

Before the interview:
1. *Check tape recorder*
2. *Bring copies of the:*
    a. *Interview protocol*
    b. *Consent form (2 copies).*
3. *Ensure that the protocol is updated with information already gathered, if appropriate.*

Informally introduce yourself and thank the participant for his or her time. Let him/her know you will be reading a formal introduction to the interview.

Thank you for participating in this interview. Before we get started, I'd like to explain the purpose of this project and request your consent to being part of this project. It is entirely your choice whether or not to participate in this study.

Project Description
This interview is part of a pilot study. This overall research project is about how different types of assessments are being used at the district, school, and classroom level, focusing specifically on middle school math. At this point we are deciding on sites for a more comprehensive study beginning next year. A second purpose is to refine our interview protocol, so please let me know if a question is confusing, if you think a question should be omitted, or if a question should be added. We are using the same protocol for many different districts, and I'll be reading all of the questions, even if we have already touched on the topic; so please let me know if I ask a question that you feel you've already answered.

During this interview you may be asked to share assessment materials. We appreciate you sharing these materials as they will be helpful resources in answering our research questions.

With your permission, the interview will be audio taped. These tapes will be used as the primary sources to refine the interview protocol and to help determine which districts we will approach for participation in the second stage of this study. The tapes will be retained until they are transcribed. Those individuals who will have access to these tapes will be the research team. Being audio taped is not a requirement for participation. You may still participate in the study should you choose not be taped.

As part of our procedures we ask participants to sign a consent form. *(Hand participant consent form).* The consent form explains the purposes of this study as well as your rights as a participant. Please take some time to read through the form and ask me any questions about your participation that you still may have. *(Check to see if participant has initialed all pages of the consent form, signed the form and how the permission-to-be-tape-recorded section is marked. Give him or her a different copy of the consent form to keep).*

(If permission is given to tape the interview)
Thank you for your time and let me know when you are ready for me to turn on the tape recorder and start the interview. *(Turn on tape recorder)*

(If permission is not given to tape the interview)
Thank you for your time and let me know when you are ready for me to start the interview.
Begin the interview:
My name is _____ and I'm interviewing _____ , who is (position) for (district).
Today is _____, and it is (time).

Interview Questions

1. PERCEPTION OF ASSESSMENT TYPES AND TERMINOLOGY

People use different terms to describe types of assessments. I would like to ask what four different assessment terms mean to you and how each type of assessment is used. The terms are formative assessment, curriculum-embedded assessment, interim assessment and benchmark assessment.

There are no right or wrong answers; we are simply interested in how the terms are used by education professionals.
a. What does the term Formative assessment mean to you?
Probe: How is formative assessment used? Do you believe this type of assessment is important? Why or why not?
b. What does the term Curriculum-embedded assessment mean to you?
Probe: How is curriculum-embedded assessment used? Do you believe this type of assessment is important? Why or why not?
c. What does the term Interim assessment mean to you?
Probe: How is interim assessment used? Do you believe this type of assessment is important? Why or why not?
d. What does the term Benchmark assessment mean to you?
Probe: How is benchmark assessment used? Do you believe this type of assessment is important? Why or why not?

2. CURRICULUM AND CURRICULUM-EMBEDDED ASSESSMENT

We are interested in the relationship between assessment and curriculum, and I would like to ask you some questions about the text series you are using for middle school mathematics.
a. Curriculum director only: QUESTIONS ABOUT THE TEXT SERIES
i. I understand that you are using _____ textbook series for middle-school math. Is this still correct
ii. Is textbook adoption done at the district or school level?
iii. How many schools are using this/these series?
iv. How many teachers are using this/these series?
v. What supplemental instructional materials, if any, do your teachers use along with the main text series?
Probes: Do you use materials from another publisher? Teacher-generated materials?
vi. Why are you supplementing the text series?
vii. Are teachers required to use the supplemental materials?
viii. How are teachers using supplemental materials?
Probe: How often do they use them?

b. CURRICULUM-EMBEDDED ASSESSMENT
Now (or "First") I would like to ask you about curriculum-embedded assessment; that is, the assessments that are part of the curriculum materials.
i. Are you familiar with the assessment materials that come with the text series you

are using for middle-school math?
ii. Would you describe the assessment materials that come with the text series you use, if any?
iii. Are teachers required to use them?
iv. How do teachers use these assessments?
Probe: Can you give specific examples of the kinds of things they learn from these assessments?
v. What do you consider to be the strengths of these assessments?
vi. What do you consider to be the weaknesses of these assessments?

c. TEACHER-GENERATED ASSESSMENTS BASED ON THE CURRICULUM
Next, I would like to discuss the assessments that middle school mathematics teachers generate on their own.
i. Do you know whether teachers are creating their own assessments based on the text series or curriculum?
ii. Are teachers required to do this?
iii. How do teachers use these assessments?
Probe: Can you give specific examples of the kinds of things they learn from these assessments?
iv. What do you consider to be the strengths of these assessments?
v. What do you consider to be the weaknesses of these assessments?

d. PROFESSIONAL DEVELOPMENT RELATED TO CURRICULUM-BASED ASSESSMENT
Now I would like to ask you about professional development for middle school math.
i. Are you aware of any professional development efforts in your district that are focused on the text series or curriculum?
If no, probe for details of any planned professional development.
ii. *Do you have any materials used in the professional development that I would be able to look at?*
iii. *Would it be possible for me to obtain or make copies of a set of these materials?*
iv. What are the content and focus of the professional development?
Probes: Does it include curriculum-based assessment? Reviewing student work? Using open-ended tasks? Extending instruction based on open-ended tasks?
v. What are the district's goals in conducting this professional development?
vi. How often do teachers participate, and for how long?

3. INTERIM OR BENCHMARK ASSESSMENT
Use this question if an interim or benchmark assessment ***is*** being used in the district. If an IA is not being used, go to the alternate form of question 3, below.
Now I would like to ask you about the interim or benchmark assessment you are using for middle school math.
a. I understand that you are using _____ assessment at _____times of year. Is this still correct?
b. *Is this assessment used by all schools and students in the district?*
**(If not)* How many schools/students are using the assessment?*

(If not) Which schools are not using the assessment?
(If not) On what basis is the decision about use made?*
c. *How long does the assessment take to administer?*
d. *Why did you choose to administer this assessment _____ times per year?*
e. Ask everyone except professional development coordinator
What other assessments, if any, did you consider?
f. Ask everyone except professional development coordinator
Why did you choose this assessment?
g. What are the district's goals in using this assessment?
h. USE OF THE ASSESSMENT
Now I would like to ask you about the ways in which this assessment is used.
i. Are you familiar with any of the ways in which the data provided by this assessment are being used?
ii. Let's start with the classroom level. How are teachers using the assessment?
Probe: Can you give specific examples of the kinds of things they learn from this assessment?
iii. How is it being used at the school level?
Probe: Can you give specific examples of the kinds of things they are learning from this assessment?
iv. And how are you using the information at the district level?
Probe: Can you give specific examples of the kinds of things you are learning from this assessment?
i. OPINION OF THE ASSESSMENT
i. What do you consider to be the strengths of this assessment?
ii. What do you consider to be the weaknesses of this assessment?
Probe: Are there other types of information you wish the assessment would provide?
j. CONTINUED USE OF THE ASSESSMENT
i. Will your district continue to use this assessment?
ii. Why (or why not)?
Probes: How/when will the decision be made to continue, discontinue or replace the assessment? Who will be involved in the decision?
k. PROFESSIONAL DEVELOPMENT RELATED TO INTERIM ASSESSMENT
Now I would like to ask you about professional development related to this assessment.
i. Are you familiar with any professional development associated with this assessment?
If no, probe for details of any planned professional development.
ii. *Do you have any materials used in the professional development that I would be able to look at?*
iii. *Would it be possible for me to obtain or make copies of a set of these materials?*
iv. What are the content and focus of the professional development?
v. What are the district's goals in conducting it?
vi. How often do teachers participate, and for how long?

3. INTERIM OR BENCHMARK ASSESSMENT
Use this question if an interim or benchmark assessment is ***not*** being used in the district
Now I would like to discuss interim and benchmark assessments.
a. Are you considering adopting an interim or benchmark assessment?

(If yes)
b. What do you hope to accomplish with an interim assessment?
c. Which systems are you considering?
d. What do you see as their relative strengths?
e. What do you see as their relative weaknesses?
f. How and when do you plan to make this decision?
Probe: Who will be involved in this decision?
(If no)
g. Why are you declining to adopt an interim or benchmark assessment?

4. FORMATIVE ASSESSMENT
Now I'd like to ask you about formative assessment for middle school math. For the purposes of this interview, I'm using the term "formative assessment" to represent the assessments teachers use in the course of their day-to-day instruction to inform their teaching. For example, this would include probing for student understanding and adjusting instruction accordingly.
a. Are you familiar with teachers' use of formative assessment for middle-school math in your district?
b. Has your district conducted any professional development focused on formative assessment?
If yes, ask questions c – h. If no, go to "Now I would like..." and start with i.
c. *Do you have any materials used in the professional development that I would be able to look at?*
d. **(If yes*) Would it be possible for me to obtain or make copies of a set of these materials?*
e. Would you describe the professional development?
i. *Who is conducting it?*
Probe: Who designed it?
ii. *How often do teachers participate, and for how long?*
f. What are the district's goals in conducting this professional development?
g. What do you consider to be the strengths of the professional development?
h. What do you consider to be the weaknesses of the professional development?

Now I would like to focus on the formative assessment your teachers are doing.
i. Are you aware of any ways in which your teachers are using formative assessment?
j. Would you tell me about the ways teachers are conducting formative assessment?
k. What information do teachers obtain from this type of assessment?
l. How do they use this information?
Probe: Please give specific examples of how formative assessment is used to guide instruction, if possible.

Ask everyone:
m. What are the challenges in implementing formative assessment in your district?
Probe: What would it take for your teachers to more fully use formative assessment?

5. COORDINATION OF PD EFFORTS (If more than one type of PD is being conducted in the district)
I'd like to ask one more question about professional development. Is there any coordination

among the various professional development efforts we have discussed?
a. *(If yes)* Please explain how these efforts are being coordinated.
Probe: Who is coordinating them?

6. Is your district conducting or planning any other professional development related to assessment that we have not covered?
If yes, probe for details of the professional development.

7. Are there any other types of assessment you are using for middle school math that we have not covered? *Probe: For example, placement tests?*
If yes, ask a, b, and c
a. Would you describe the assessment?
b. What information do teachers obtain from this assessment?
c. How do they use this information?
Probe: Can you give specific examples of the kinds of things they are learning from this assessment?

Thank-you, those are all the questions I have. Do you have anything you'd like to add?

8. *Close the interview and thank the participant again. Ask if it is OK to call or email with brief follow-up questions, should they occur to you later. Make a note of the best phone number or email address to use.*

Appendix B:
A Summary of Perie, Marion, and Gong's (2009) Framework

Table B1

A Summary of Perie, Marion, and Gong's (2009) Framework for the Purposes of Interim Assessment

Evaluative	Instructional	Predictive
Primary goal		
Programmatic assessment designed to provide evaluative information in order to change curriculum or instruction not necessarily in mid term but over the years.	Adapt instruction and curriculum to better meet student needs and meet the learning goals.	Determine each student's likelihood of meeting some criterion score on the end-of year tests (such as state tests or exit exams) or success with postsecondary curriculum.
Examples		
Provide aggregate information on student achievement at a district level.	Evaluate how well the student has learned the material taught to date.	Predict students' performance on a summative assessment.
Enforce some minimum quality through the standardization of curriculum.	Provide a more thorough analysis of the depth of students' understanding.	Determine whether students are on track to succeed on the summative assessment.
Ensure that teachers are staying on track in terms of teaching the curriculum in a timely manner (i.e., pacing).	Provide specific feedback on where the gaps in a particular student's knowledge are at the classroom level.	Diagnose and provide corrective feedback to help a group of students get on track to succeed on the summative assessment.
Provide information to help the instructor better teach the next group of students by evaluating the instruction, curriculum, and pedagogy.	Motivate and provide feedback to students about their learning.	
Determine whether one pedagogical approach is more effective in teaching the material than another.	Determine whether students are prepared to move on to the next instructional unit.	

Appendix C:
Substantive Codes

Purposes:

Educator Evaluation (Evaluative)

Implementation:

Assist with Data-Driven Dialogue

Engage in Teacher Collaboration

Use of Protocol

Inform PD or Resources

PD on Informing Instruction

PD on User Training

Little/No PD Received

Require Compliance

Practices/Uses:

Feedback (Instructional)

Differentiation/Grouping (Instructional)

Placement (Instructional)