Educ Asse Eval Acc (2010) 22:29–52
DOI 10.1007/s11092-010-9090-2

Formative assessment systems: evaluating the fit between school districts' needs and assessment systems' characteristics

Matthew Militello · Jason Schweid · Stephen G. Sireci

Received: 9 September 2009 / Accepted: 7 January 2010 /
Published online: 27 January 2010

Abstract Recent legislative and local school accountability efforts have placed a premium on the collection, analysis, and use of student assessment data for educational improvement. As a result, schools have sought assessments that will provide additional information about student performance. In response, a burgeoning boon industry formed—formative educational assessment systems. In this study we describe how districts search for and acquire formative assessment systems to meet their needs. We focus on three school districts that adopted three different formative assessment systems. Our findings suggest the fit between a system's characteristics and a school district's intended use is the most important consideration in instituting a successful formative assessment system that will have a positive impact on teacher education and student learning.

Keywords Formative assessment · Case study · Assessment system characteristics · Assessment use

1 Introduction

Educational accountability in the United States has traversed a wide path from Horace Mann's calls for uniformity of the one-room schoolhouse to the development

M. Militello (✉)
North Carolina State University, 608N Poe Hall, CB 7801, Raleigh, NC 27695, USA
e-mail: matt_militello@ncsu.edu

J. Schweid · S. G. Sireci
University of Massachusetts at Amherst, Amherst, MA, USA

J. Schweid
e-mail: jschweid@educ.umass.edu

S. G. Sireci
e-mail: sireci@acad.umass.edu

of an industrial workforce to the race to space to the current preparedness for a global economy. With each new wave of accountability, school districts had to enact new reform mandates (e.g., Carnegie Units, vocational courses, math, science, and foreign language requirements, etc.). As the United States entered the 21st century new federal legislation (No Child Left Behind) replaced the prior reform focus of mandated strategy implementation to targeted student achievement outcomes. As a result, educational accountability today is synonymous with student achievement outcome testing and the sanctions that accompany the results (Darling-Hammond 2004; McDonnell 2004; Ogawa et al. 2003). States have moved forward to create approved state-level assessments and targets for districts, schools, and students to make adequate yearly progress (AYP). In turn, states have pressed individual school districts to meet student achievement targets on the state assessments.

Early district efforts where characterized by schools aligning (or re-aligning) their curriculum to match what was taught on the state assessment. However, alignment is only one step toward improving student achievement. The U.S. Department of Education (2003) stated, "Research shows that teachers who use student test performance to guide and improve teaching are more effective than teachers who do not use such information" (p. 2). However, many educators find the collection, interpretation, and use of school data difficult (see Coburn and Talbert 2006; Militello 2004; Ogawa and Collom 2000; Petrides and Guiney 2002; Wayman and Stringfield 2006). The data they have available do not always meet their needs. For example, classroom-level assessment data are often not a valid measure of student achievement as it relates to the state-level assessment. Additionally, classroom assessments are not systemically implemented which limits the dialogue or professional development possibilities. The data school receive are usually normative and summative in nature; lack strong curricular links (e.g., Stanford, ITBS, ACT, SAT, state-level assessment, etc.), and arrive too late to be useful for instruction. That is, assessments schools have access to do not provide timely, diagnostic-level data that could be used within school year for educators. As a result, schools have been characterized as in "the paradoxical situation of being both data rich and information poor" (Wayman and Stringfield 2006, p. 464). Increasingly, educators and researchers have questioned the utility of currently available data in schools and have realized different data are needed to change their pedagogy and decision-making to improve student achievement (Baker 2007; Popham 2004, 2008). This has led districts to search for an assessment system that can inform their work.

Not to miss an opportunity, numerous formative assessment products have emerged on the educational scene. Formative Assessment Systems (FAS) are a fast-growing and under-studied phenomenon, with major implications for educator practice. FAS packages typically include electronic databases, access to item banks with pre-aligned questions tailored to the state's curriculum frameworks, online and/or paper exams that are adaptive in some cases, automated analytic functions, easy to read graphic reporting mechanisms, and technical support. These systems offer the potential to close the "virtuous circle" of curriculum, instruction, and assessment by providing "just-in-time" feedback for teachers and administrators (Popham 2004, 2008). However, a question remains: will formative assessment systems fulfill the call for more meaningful information than currently exists? Moreover, given the

diversity of the characteristics among formative assessment systems, understanding the relationship between each system and its intended use by a district is crucial.

This article draws on school district reaction to assessment accountability to investigate the "fit" between the intended purpose and system characteristics in three Northeastern school districts using three different formative assessment systems. Our analysis uses data from interviews with educators at multiple levels of each district as well as the technical features of each assessment system. Our findings suggest that the districts did not account for the fit between the intended use and the characteristics of each formative assessment system. Each case provides an account of how assessment data can be used in districts based on the alignment or fit between a district's purpose and an assessment system's characteristics.

2 Related literature

The press for more useable information on student achievement has led to the creation of a new booming Formative Assessment Systems (FAS) industry. FAS combine tightly knit assessment instruments, data-warehousing, analysis, and reporting batteries (Halverson et al. 2007b; Sharkley and Murnane 2006). In the past five years over 20 different companies have developed and marketed different forms of FAS (Militello et al. 2008). These systems offer the hope of more appropriately tailored sources of information. Demand for formative assessments has skyrocketed, and schools are deploying a variety of models, ranging from "home-grown" tests created by teachers themselves to commercially packaged assessment systems costing $12 or more per student. For many, however, the fear exists that assessment companies and even district based assessment specialists, looking to capitalize on both the need for information and relative lack of knowledge about type of information needed to actually inform classroom decisions, have simply repacked non-specific assessments and sold them as formative, decision making tools (Baker 2007; Popham 2008). Thus, there is a critical need to assess the fit between schools intended assessment uses and the characteristics of a given assessment (Militello and Heffernan 2009).

The notion of utility has become a contentious issue in the varied conceptions of test validity (Messick 1989). Here, the underlying judgment becomes one of coherence; is the assessment instrument in question useful in providing information tailored to a specific set of user defined purposes? The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education 1999) underscores the importance of fit between assessment use and purpose by defining validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). Similarly, Kane (1992, 2006) posits that an evaluation of test validity involves evaluating the degree to which the intended uses of an assessment fit its characteristics, and gathering evidence to support the use of a test for a particular purpose. Consequently, evaluating the fit between districts' intended uses of formative assessments and the technical characteristics of the assessments is needed to justify the utility of a FAS.

2.1 Districts' intentions

A growing body of literature highlights the diversity of purposes of FAS in schools today. The purposes of creating or implementing assessment systems in schools can be summarized as assessment data that: Increases reliability and predictive validity of school improvement efforts (Coburn and Talbert 2006), helps improve instructional practice (Coburn and Talbert 2006; Militello 2005; Murnane et al. 2005; Wayman and Stringfield 2006), initiates conversation and pedagogical collaboration (Lachat and Smith 2005; Wayman and Cho 2009), develops a sense of ownership for achievement (Halverson 2003; Militello and Sykes 2006; Murnane et al. 2005; Weiss 1995), reveals curricular adherence and alters alignment (Coburn and Talbert 2006; Kerr et al. 2006; Murnane et al. 2005), monitors teacher effectiveness (Militello 2004; Murnane et al. 2005), guides long-term, district-wide planning (Brunner et al. 2005; Coburn and Talbert 2006; Gitomer and Duschl 2007; Militello 2005), targets student needs for short-term intervention (Halverson et al. 2007a; Militello 2004; Murnane et al. 2005), and assesses readiness for state-wide summative assessment (Murnane et al. 2005; Streifer and Shumann 2005).

Not surprisingly there are a number of stated purposes for assessment data in schools today. Diversity of thought from different actors in various levels of an organization is nothing new. What the multiple purposes call for is a better understanding of characteristics of different assessment systems. Before describing our study, we first describe the characteristics of educational assessments since these characteristics can differ across various assessment systems.

2.2 Evaluating and understanding characteristics of formative assessments

Evaluating the appropriateness of formative assessments involves evaluating the technical quality of the assessments and the degree to which the assessments provide the desired types of information. As the *Standards for Educational and Psychological Testing* (AERA et al. 1999) and validity theorists tell us (e.g., Kane 2006; Messick 1989) a sufficient body of evidence is needed to justify the use of assessments for specific purposes. The *Standards* categorize validity evidence into five sources—evidence based on (a) test content, (b) response processes, (c) relations to other variables, (d) internal structure, and (e) consequences of testing. Other psychometric properties of tests, such as adequate score reliability, and consistency of score scales over time (e.g., when tests are used to track growth) are also important for evaluating test utility.

Currently, educators understand validity evidence in terms of the alignment of test content with valid academic outcomes and the degree to which test results provide insights into the material students have or have not mastered, as well as their thinking and reasoning processes (Coburn and Talbert 2006). However, little research exists that illuminates the actual format and psychometric characteristics of current assessment systems and the instruments they use. In fact, most data on system characteristics resides in proprietary technical manuals and commissioned studies by assessment companies.

Large scale summative achievement measures, both criterion-and norm-referenced, are bound by validity and reliability standards (AERA et al. 1999).

These assessments have garnered the lions share of funding, attention, and place within the current educational paradigm (Pellegrino et al. 2001). When large amounts of state funds are available for developing large-scale assessments, such as those used to evaluate adequate yearly progress under No Child Left Behind (NCLB) states can require their testing vendors to provide technical documentation that supports the use of the test for Adequate Yearly Progress (AYP) and summative assessment purposes. However, less funding and control are typically available when districts delve into formative assessments. Formative assessments are often administered periodically throughout the school year and assess a variety of concepts for use at a variety of school levels (Militello et al. 2008; Perie et al. 2007). These tests blend aspects of both authentic, classroom assessments with large-scale test principles. A major purpose for intermediate assessment is the measurement or inference of academic progress (Datnow et al. 2007). These tests (AKA interim or benchmark assessments) are normally aligned to a set of standards. These standards range from national curricular models to state-based curricular benchmarks to local district curricular standards. In general, these products claim to provide information about student attainment of curricular benchmarks and standards, student growth between assessment cycles, prediction of state assessment score, and teacher curricular adherence. The utility is dependent not only on the curricular standards they test, but also to test, item characteristics, and other information that would support their utility (Baker 2007).

Today, formative assessments are understood to be activities that inform teachers so they can modify their daily instructional practice (Black and Wiliam 1998a, b; Wiliam 2006). However, not all formative assessments are the same. A prominent distinction lies in assessments *of* learning and assessments *for* learning (Black et al. 2003; Stiggins 2005). Assessing *for* learning stipulates that assessment data in at the diagnostic-level and timely to enable educators to use the data in their practice and decision-making. Here the definition of formative assessment becomes more delineated by utility for teachers. That is, do the data provide teachers with details about if and how students are learning? This type of formative assessment has been called cognitive diagnostic. Leighton and Gierl (2007) defined cognitive diagnostic assessment as a means to "measure specific knowledge structures and processing skills in students so as to provide information about their cognitive strengths and weaknesses" (p. 3). Popham (2008) expands on this definition by explaining the importance of understanding the assessment as part of a greater process of adjustment:

> Formative assessment is a planned process in which assessment-elicited evidence of students' status is used by teachers to adjust their ongoing instructional procedures or by students to adjust their current learning tactics. (p. 6)

Shepard (2000) explained there are several tenets of classroom assessment that differentiate it from past models of summative assessment and measurement theory: that it is iterative and ongoing; that it has an intentional focus on transfer of understanding rather than merely memorization; that it is a form of self-assessment; and, importantly, that it acts not just as a method of evaluating learning but teaching as well. As a result, in the race to create and implement formative assessments, the important analysis of creating a typology of formative assessment systems has been widely ignored.

2.3 Summary of related literature

There is literature on the purposes of assessment systems and the characteristics of these systems. This research, however, has been isolated. That is, few studies have investigated the fit between schools' intended purposes and the characteristics of an assessment system. Thus, there exists an unrefined and immature understanding of the relationship of system characteristics and user intent. The purpose of this study was to understand both the rationale and intent to the district's FAS selection and the characteristics of each system. Our overarching question focused on the *fit* between the system and its intended use: *To what extent does the fit between intended use and system characteristics foster or inhibit the ultimate utility of formative assessment systems for schools?* To determine this we focused on three specific sub-questions: (1) Intended Use: *What data and action did each district want from the assessment system?* (2) System Characteristics: *What were each of the formative assessment systems designed to do?* (3) Actual Use*: How are school district educators using the assessment systems?*

3 Research design

The data summarized in this article come from a year and half long study involving three northeastern school districts that employ formative assessment systems. To choose districts, we first identified all prominent formative assessment companies and the contracts they held in a Northeastern state. We then consulted with the state department of education to identify specific districts that exhibited prominent use of formative assessments. We narrowed our selection based on the implementation on the FAS in middle school mathematics. After conducting phone interviews with state department of education assessment personnel the three different school districts were selected. Each district was contacted and volunteered to participate in the study. The three districts involved include Ryder Public School, Woodbury Public Schools, and Franklin Public Schools (we use pseudonyms for all schools and names). Within each district our research focused on one specific school and their experience implementing and using a FAS.

3.1 Participants

Ryder Public Schools (RPS), located within of major metropolitan area, is relatively small, serving eleven schools that serve approximately 5,839 students. The middle school we investigated, Ryder Middle School (RMS), is one of three middle schools in the district. RMS has an enrollment of 526 students (41% White, 29% Hispanic, 14% Asian, 3% African-American) in grades 6–8. Seventy-three percent of the students qualify for free or reduced lunch. In 2004 Ryder Public Schools conducted an audit the district's academic practices and concluded, among other things, that it would benefit from: (1) a district-wide curricular audit (including the development of a scope and sequence, "power standards," and pacing guides all of which would be synchronized with the state curricular benchmarks and standards) and (2) the creation and use of a district-wide formative assessment system. After investigating a

number of formative assessment systems, the district created and implemented an "in-house" formative assessment system.

Woodbury Public Schools (WPS) is an urban school district that is diverse both ethnically and economically. The district serves just over 23,000 students. Woodbury Middle School (WMS) is one of four middle schools in the district. The school serves over 900 seventh and eighth graders (54% White, 23% Hispanic, 7% Asian, 7% African-American). Fifty-three percent of the students qualify for free or reduced lunch. In 2002 staff at Woodbury Public Schools had grown disappointed with the state assessment. In 2003, the school district purchased and implemented a formative assessment product: Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP).

Franklin Public Schools (FPS) straddles the line between a suburban and urban setting. FPS enrolls 5521 students. Franklin Middle School (FMS) is one of four district middle schools, located in the historic city center, houses 441 students in grades 5–8 (47% White, 31% Hispanic, 11% Asian, 11% African-American). Sixty-three percent of the students qualify for free or reduced lunch. In 2003 the district hired a new assistant superintendent who had implemented a state data analysis system in her past district. In 2004 the district received a grant from the state department of education to fund the Assessment Technology Incorporated's (ATI) Galileo formative assessment system.

3.2 Procedures

Data were collected in three stages throughout the 2006–2007 school year. The first stage focused on district-level administrators who helped oversee and implement assessment efforts. This included superintendents and assistant superintendents ($n=7$) and assessment coordinators ($n=3$). The interview protocols at the district level focused largely on the organizational history and goals pertaining to formative assessment. More specifically the main themes of the district level protocol included: (a) the reasons why a formative assessment system was viewed as necessary, (b) the districts search for a specific formative assessment systems, (c) the specific rationale for choosing a formative assessment system, and (d) the desired use of chosen formative assessment system. We also attended district level data meetings and collected assessment related artifacts (e.g., professional development opportunities and internal memos).

The next stage involved the investigation of each specific formative assessment system. Here we draw from interviews with test developers (psychometric staff at NWEA and ATI and district developers in the Ryder Public Schools), technical documentation from NWEA and ATI, and attendance at a formal presentation and question and answer session with representatives from NWEA and ATI. The interviews focused on understanding the stated purposes of the assessments and the degree to which evidence was available to support the use of the assessments for those purposes.

School level educators were the focus of the third and final stage of data collection. At the school level, data were collected from interviews with principals, guidance counselors, department chairs, math coaches, teachers, as well as observational notes from math departmental meetings and grade level data meetings. Interviews included principals ($n=3$), counselors ($n=4$), curricular coaches and department chairs ($n=3$), and math teachers ($n=21$). At the school level interview protocol focused on educators' behavior using formative assessment data.

3.3 Data analysis

The analysis of the data collected can be characterized by a coding phase and comparative analysis. The first stage involved memo creation from field notes, artifacts, and interview transcriptions (Merriam 1988). The memos and transcribed interviews were then exported into a computer database. Memos and transcripts were then described using an open coding system (Bogdan and Biklen 1992; Miles and Huberman 1994). Codes were used that signified themes of use and purpose previously identified in extant literature. Additionally codes were developed for findings that lay outside reported FAS understandings. Simultaneously, members of the research team with a background in assessment validation reviewed collected technical information to interpret system validity and utility. Here technical characteristics of were condensed and described using similar coding system. The final stage of data analysis employed the two sets of coded data using a cross-comparative strategy to understand the alignment of the identified phenomenon, conditions, contexts, consequences, and intervening conditions (Creswell 1998; Strauss and Corbin 1990). This analytical technique helped the researchers find patterns and build explanations for the gaps identified between the intended use and the actualized use of assessment data in each district.

4 Results

We summarize the results of our study by first describing the reasons why each district wanted a FAS and its expected outcomes. Next, we describe the system characteristics of each FAS. Third, we describe preliminary findings regarding how each district is using their formative assessment system.

4.1 How each district intended to use a formative assessment system

Each district had a different process for coming to understand their needs for a formative assessment system. For Ryder Public Schools three elements were important: (1) The assessment would evaluate student performance, (2) The assessment would assist teachers in planning future curriculum, and (3) The assessment would have predictive qualities for the annual state assessment system. In other words, the desired assessment would assess student mastery of taught material, explain the current student competencies of untaught materials, and provide information as to how students may fare on the state. The district director of math explained, “We decided that the assessments should be holistic-covering all of our standards- so teachers have some data that will drive their instruction, in terms of what they taught and what they were planning to teach.” The director went on to say, “We wanted teachers to start talking more about student work, so by having these assessments that the teachers worked on it represented a common goal.” As a result, the district math director formed a committee to design a district-based assessment system.

Woodbury Public School wanted an assessment that would provide a detailed diagnosis of student growth. Additionally, WPS wanted an assessment that would

provide them with timely-*in-school year*- results. For a district the size of WPS this had always been a challenge. As a central office administrator explained, "One of the things for us is our mobility rate. This gave us a uniform thing to check how students were doing even if they changed schools in our district. Also, new kids coming in, we didn't have to wait until we could give kids a test and figure out where they were. We needed an assessment to ascertain students' ability immediately." WPS wanted an assessment that would provide baseline data for each student and to track growth of students. Consequently, the district began to investigate assessments that would provide the kind of growth data they wanted. WPS decided to invest in a commercial product, the Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP) testing.

In 2004 the Franklin Public Schools leadership team made the decision to implement a formative assessment system. The technology/curriculum specialist described the process: "We started to look at the absence of usable data and we started to realize that what we did have didn't seem to have a correlation with state. One of the first things we said was, 'Oh my God, we can't wait until the end of the year to determine whether we had been successful or not', we needed to have a way of actually differentiating which students needed extra help." The assistant superintendent added, "We wanted to know where we needed to make interventions." Consequently, the district wanted an assessment tool that would provide teachers with student-level diagnostic data. As the assistant superintendent explained, "We want to move beyond teaching for compliance… [toward] a clear focus on student learning." After an intensive and inclusive review of formative assessment systems, FPS decided upon ATI's Galileo system. In their view the Galileo formative assessment system would fulfill their intended use of assessments: (1) assess learning standards-based learning by student, (2) generate conversations and actions around student learning and instruction, (3) provide the district with *just-in-time* data, and (4) provide the district with a data warehouse in order to make multiple sources of school data accessible.

Each district wanted an assessment of the taught curriculum. However, each district differed on other intended uses. Ryder also sought a predictive instrument for the state assessment, Franklin focused on student diagnostic information, and Woodbury wanted student-growth data. Next we describe the characteristics of each of the formative assessment systems selected.

4.2 Characteristics of each formative assessment system

In this section we describe each of the formative assessment systems. We describe the technical features of each assessment including: score reporting, content validity and alignment, and data warehousing capacities.

4.2.1 Ryder's district developed assessment system

The Ryder Public School assessment system is called the "Quarterly Assessment." A district curriculum group (the district math director and a number of district math teachers) created the math quarterly assessments (QA). The math QA was piloted in the 2005–06 school year and fully implemented across multiple grades in 2006–07.

The questions for each assessment came from two sources: the pool of state assessment released items and those developed by the committee. No other technical work was done during the assessment development process at Ryder. Each QA consists of 30 multiple-choice questions that cover all of the Ryder math standards (benchmarks). The multiple-choice portion of the assessment is broken into two sets of questions: (1) previously taught benchmarks and (2) not yet introduced benchmarks. For example, the November QA consisted of seven multiple-choice questions from the material recently taught and the remaining questions based on the standards from the next three marking periods. The intent of this design was to gain an understanding of student mastery, what *was* taught, and of general student knowledge, what was *to be* taught. Additionally, each assessment has four short answer questions and one open response question.

The QA are scored and loaded in a Microsoft Excel spreadsheet by the district math director. The schools administer the assessments and send results to the math director on a Friday. The math director then works over the weekend to provide the results to schools and teachers the following Monday. Assessment data are then disaggregated to the class level. That is, schools and teachers receive a report that provides results for the overall class by standard. Results are not disaggregated by item for individual students or by subgroups of students—although teachers could go through each Scantron sheet themselves. In addition to class level reports for teachers, building administrators are provided school wide reports. Both reports generate one type of graph: percent of students answering each item correctly.

4.2.2 NWEA's MAP

The MAP Math test consists of 52 multiple-choice items. Fifty of those items are used to calculate scores for students; the other two are tryout items that are being piloted for future use. There are five response options for each multiple-choice item on the Math test. For the Reading test, there are 42 items—40 scored items and two tryout items. The Reading items have four response options. At least 7 items within each goal area are administered to each student to ensure the goal scores have sufficient precision. According to the NWEA *Technical Manual* (2005), there are at least 1,500 items in the item pool for each subject area tested in a state with a minimum of 200 items per goal area. MAP tests are supposed to be un-timed, but the manual recommends allotting 75 min to ensure all students will finish without being rushed. A one-month testing window is necessary because each student needs access to a computer terminal.

There are several features of the NWEA assessments that need to be described first to understand their strengths and limitations. First, MAP assessments are drawn from very large item pools that span grade levels. The items within the pool are coded according to various content characteristics at both the global content strand level and at more specific (objective) levels. By defining heterogeneous item pools that span grade levels in each subject area, the tests can target specific content using items of various difficulty levels. A second key feature of MAP is its use of *computerized-adaptive* testing technology. This assigns test items to students by matching the difficulty of an item to the achievement level of the student. The content being measured by each item is also taken into account. A third key feature

is that all items within each subject area pool are calibrated onto a common scale. This scaling allows students to be placed onto this same scale even though they respond to different items. The common scale also allows for analysis of student growth across time. A fourth key feature of MAP is the customization of the item pools in each subject area for each state (or district, if desired). State-specific item pools are created from the "universe" of all MAP items so that the pools used for a particular state are best matched to the state's curriculum frameworks. The end result of this system is the administration of tests that: (a) are aligned to state curricula; (b) are targeted to students' proficiencies in a way that minimizes testing time, but maximizes measurement of academic growth; and (c) allow for comparisons of students' performance to national norms. Items within NWEA item banks are written to be as widely applicable as possible for measuring specific objectives.

All MAP items are calibrated using a one-parameter item response theory (IRT) model. The single parameter for each item is a difficulty parameter, which represents the location of the item at the point on the IRT scale where students have a 50% chance of correctly answering it (see Hambleton et al. 1991 for a complete description of IRT models). NWEA transforms the Rasch IRT scale to a RIT scale on which all items are calibrated and students' scores are reported. This RIT scale extends from elementary through high school and ranges from approximately 150 to 300 with a mean of 200 and a standard deviation range of about 12–15 points. This allows school educators to chart growth by student over time.

Students' MAP scores are reported in several ways. Test results for individual students include a RIT score range, percentile score range, and goal performance classifications for each goal area. The RIT score and percentile ranges give the student's RIT score and percentile rank alongside the lower and upper bounds of the 68% confidence interval surrounding those scores. These ranges reflect the likely RIT scores or percentile rank the student would earn if they took the MAP again without any additional instruction in the subject area. For the reading test, "Lexile" scores may also be provided. These scores can be used to select reading material appropriate for the student's current reading proficiency. This information can be used to identify strengths and weaknesses for students, relative to national norms. NWEA also uses the 40th percentile rankings (from the national norm group, not by specific state) to indicate projected performance on state-level assessments.

MAP results are also reported in various aggregated ways to summarize student performance at the classroom, school, or district level. For most reports, teachers and administrators can log onto NWEA's website through a secure portal, and view school-wide and classroom reports broken down by whatever variables are of interest (e.g., growth, proficiency, and growth targets for each student and indices indicating distance from targets). Other reports are available, too, through the "Descartes" system, which links students' test results to NWEA's "Continuum of Learning." Descartes can be used to identify appropriate instructional material for individual students or groups of students with similar performance on MAP tests.

4.2.3 ATI's Galileo

Galileo is a system for building benchmark assessments available through Assessment Technology Incorporated (ATI). The system involves a large bank of

items from which unique assessments are developed to best suit the needs of a particular school district. The Galileo system is designed to enable ATI to work collaboratively with a district to design an assessment system that is aligned with local instruction and informs curriculum planning. As described by Bergan et al. (2006), "Benchmark assessments are locally relevant, district-wide assessments designed to measure student achievement of standards for the primary purpose of informing instruction" (p. 3).

Galileo offers two forms of assessments: a district-level benchmark assessment and an individual teacher-created formative assessment. The benchmark assessments were used in FPS. ATI begins by working with a district to identify the number of benchmark assessments to be administered in a given year and the standards (objectives) to be measured at each subject area in each grade level. Districts are able to custom order assessments through Galileo's "Educational Management System." ATI has an online Benchmark Planner in which the district defines the assessment goals, specifies the standards to be measured and the number of items per standard, and reviews preliminary versions of the assessments.

Galileo has several reports that summarize students' performance at the individual and aggregate levels. Aggregate level reports can be produced at the class, school, and district level, and many can be created interactively to suit the user's needs. A primary report for the benchmark assessments is the Developmental Profile Report. This report lists and describes all the standards (objectives) measured and provides an achievement level classification for each student for each standard.

ATI determines the cutoffs for the achievement level classifications by asking the district how many achievement levels they want, and then divide the number of items measuring the standard by the number of achievement levels. For example, if there are nine items measuring a standard and the district requests three achievement levels, students who get 7–9 items correct will be classified as "meets standard," students who get 4–6 items correct will be classified as "approaches standard," and students who get less than four items correct will be classified as "falls below standard." This process used to set these standard-specific achievement classifications is criterion-referenced, but it is essentially arbitrary and does not take into account the difficulty of the items.

Achievement reports range from the classifications generated for aggregate scores to item level analysis. For the aggregate reports, the percentages of students in each achievement level classification are presented for each standard. Teachers can also look at a Class Development Profile Grid Report, which lists the achievement level classifications for each standard for all students within a class. Item Analysis Reports are comprehensive and present information regarding students' performance on all items on a benchmark assessment. For each item, the standard being measured is listed as well as the percentages of students who selected each response option (i.e., the percentage choosing the correct item and each incorrect response option). Those percentages are stratified by five percentile rank score intervals, so that the teachers can see the response options chosen by students of different achievement levels. By comparing this statistical information with the test items, teachers can see the types of mistakes their students are making.

Galileo also produces a Risk Assessment report as part of its Aggregate Multi-test report. This Risk Assessment classifies students as "high risk", "moderate risk",

"low risk" or "on course" with respect to not meeting the state's proficiency requirement in a subject area. ATI typically sets the cut scores for this risk assessment using equipercentile equating. In this process, the percentages of students on the statewide test who are classified into a specific achievement level or higher (e.g., "needs improvement" or "proficient") are noted and then the Benchmark cut score is set at the point that corresponds to the same percentage on the Galileo Benchmark assessment (Bergan et al. 2006). These Risk Assessment reports can be aggregated up to the district level or disaggregated down to the student level.

4.2.4 Summary of assessment system characteristics

Table 1 summarizes our findings of the technical characteristics of each formative assessment system. As the brief comparison indicates, all systems reported results quickly, but differed with respect to the types of information they provided and their technical characteristics.

4.3 Use of each formative assessment system

The third phase of this study focused on how the formative assessment systems in each district were being used. In this section we provide an overview of how the systems were implemented, and we report on how the assessments were being used.

4.3.1 Ryder's use of a district-developed FAS

In RPS, the math QA is administered in grades six through 12 five times per year. The math assessments are administered in September (pre-test), November (QA), January (QA), March, (QA), and June (QA), coinciding with the end of each marking period. The assessment window is usually two to three weeks prior to the end of the marking period. Schools are sent the assessments and asked to administer them and send in the results by Friday of that week. Each test is allotted two 47-minute class periods for completion. Each school decides on the best way to implement the assessment.

Class level data are used frequently as a jumping off point for teacher discussion. A teacher described the purpose of such discussions as follows: "The district uses the assessments as kind of seeing what we did teach and what we didn't." Other

Table 1 Summary of formative assessment system technical features

Feature	District-developed (Ryder)	NWEA's MAP (Woodbury)	ATI's Galileo (Franklin)
Score Reporting (Time & Access)	Quick (48 hrs./relies on district personnel)	Quick-24 hrs. (teacher access)	Quick-24 hrs. (teacher access)
Content Validity/ Alignment	Excellent	Moderate	Excellent
Data-Warehousing	Insufficient: QA data only in Excel spreadsheets	Moderate: MAP Data	Excellent: Galileo, state assessment, and local data (e.g., open responses)

teachers echoed this sentiment. On each of the Quarterly Assessments there are 30 questions, only seven of which directly pertain to content explicitly covered during that period of time. That is, each assessment only has a small set of questions on the content that was most recently taught. The remaining questions are drawn from either content taught in previous marking periods or in future periods. In reference to using the data on content not yet covered one teacher joked, "You can draw a line down your chart and ignore everything to the right." As a result, we found that the questions about content and standard that were most recently taught yielded the most fruitful information to inform teachers' practice.

Some teachers believed the greatest utility of the Quarterly Assessment was in planning and curricular decision-making. One teacher stated, "There are some questions that our students do really well on, so we don't really need to focus on it as much." Other teachers used these tests to identify areas of weakness and strength to alter their scope and sequence. However, most teachers agreed that there was little opportunity to actually go back and revisit material if the test indicated a drop in performance. For example, teachers told us that even if the results indicated that student were not learning the material, there was no time to re-teach or implement specific pedagogical changes (e.g., differentiated instruction, re-grouping). As one teacher explained, "In theory we have the time to go back and teach but there is no time. There is pressure to move on to the next part of the curriculum."

Ultimately, most instructional decisions in RMS classrooms were made using teacher created assessments or the state assessment. One teacher explained differentiation choices as follows, "We all use regular in-class assessments for grouping." When asked about the lack of student-level data from Quarterly Assessments, another teacher explained "We know our students so we would know who do poorly, or need more help." This sentiment has led many teachers to believe the form of the assessment may not be best geared to understanding their students. As one teacher explained, "I think if we tested what we preciously taught by quarter we would get more useful information." Another teacher emphasized this sentiment by adding "Sometimes it's frustrating to see our kids bomb on questions we did not even teach yet."

4.3.2 Woodbury's use of NWEA's MAP FAS

Woodbury Public Schools (WPS) uses the MAP math assessment in grades 2 through 10 four times a year. School educators have been charged to use the data provided by MAP for curricular changes, grouping decisions, and oversight. Teachers at WMS indicated they have engaged in looking at MAP results both independently and in groups to identify patterns in student scoring. During debriefing sessions with their department chairs, math teachers identify weak areas for their classes. Some teachers then devise ways to incorporate weakness areas into their district-mandated content. When pressed, teachers admitted to not having detailed enough information to really hone in on specific areas for focus. As one math teacher explained, "to some degree you're winging it, but at least you are in the ballpark." The department chair explained that he encourages teachers to use the MAP suggestions for their lesson planning. Additionally, many teachers assign students to computer-based remediation and supplementation based on MAP results. A principal explained, "They are talking about best-practice, new techniques… [MAP results] are very informative for this."

Consequently, MAP data have begun to initiate conversations about not only student achievement, but also teaching practices.

Educators at WMS report that students have embraced the MAP testing process and the feedback they receive. Teachers in WPS have increasingly been using MAP scores to help students set performance goals. After each test, teachers conference with each student basis and review their scores. At that point teachers have students develop score goals for the following administration. This process is not formal, there are no templates or reporting forms, but some teachers do take the time to record student goals. With the aid of MAP reports, teachers and students can easily track progress across the school year. Additionally, the schools use the MAP Lexile ratings to provide parents with valuable information about their children's reading performance. Both the public and school library systems in Woodbury are cross-cataloged by Lexile score. Thus, students, parents, and teachers have a common language of understanding specific reading abilities of each student. However, teachers remained skeptical of results as they are not provided access to the actual questions that were administered to their students. They felt that testing was too frequent and the data results they received were not helpful. One teacher explained, "I am not getting all I can out of MAPS. In fact, I want more specific data about where my students are having trouble."

In the end, the NWEA-MAP was intended to understand specific measures of growth for students. As a central office administrator explained, "The test really provides us a series of snapshots of how kids are doing." Another central office administrator went further, "I think that test in a lot of ways is a validation—the teacher had kids in their class all semester… this information is about the kids in your class with the results that you can do something about." However, building-level educators questioned the viability of using the MAP assessment data to integrate specific pedagogical practices.

4.3.3 Franklin's use of ATI's Galileo FAS

FPS administers the math Galileo benchmark assessment three times each year in grades five through eight. Tests are administered in the beginning of November, the end of January, and at mid-April. Each test consists of 35–40 multiple-choice items. For each assessment, between six and eight standards are covered, two of which are standards that had previously been flagged by the district for retesting. Additionally, FPS has opted to add open-ended and short answer questions to each benchmark test. Although ATI provides an on-line administration option, FPS administers Galileo in paper-and-pencil format using Scantron answer sheets (due to costs in terms of money, access to computers, and people hours). Once the assessment is completed, trained school officials "scan" each score sheet and email the results to ATI. All coding of student information and Scantron forms are provided by ATI.

FPS has developed a support process for the use of the Galileo data that includes a number of "trigger" mechanisms. That is, when the data are reported the results are followed-up with a set of standard operating procedures. The main forum to discuss the Galileo data resides in debriefing sessions with math teachers. The math curriculum specialist explained,

> While we talk at district level about what standards are being missed, when we talk to teachers and they have their student reports we are able to ask more

specific questions like: Why didn't the kids get this? We ask what are you thinking about this? Why did your kids think about this? What are you going to do in the class? It's about getting it back into the classroom.

The debriefings position teachers to use assessment information to understand learning more than performance. As explained by the technology/curriculum specialist, "We are changing the idea of assessment *of* learning to assessment *for* learning."

The debriefing sessions revolve around three overarching questions: (1) How well did our students perform on the assessment (as a school and individual students)? (2) Why is this occurring? and (3) What can we do about it? Grade level math teacher groups discuss item-level analysis of their student performance data. As the curriculum specialist explained, "powerful conversations are the rule not the exception. We have developed a data-rich culture in the building."

These discussions have led teachers and administrators in FPS to use Galileo testing to drive curriculum changes, interventions, and even student participation. For example, educators are regularly engaged in discussions (in debriefing sessions and on their own) to identify students who scored poorly and how to modify instruction. As one teacher explained, "We look at students who fell short… We add questions based on this information." Most sessions begin by looking at questions that less than 70% of students answered correctly. The items are analyzed and discussions are held. These discussions are reported on a data analysis template provided by the district staff. These forms include not only analysis of data, but provide teachers with the opportunity to create a specific plan of action based on the analysis. .

Teachers have also begun to find ways to include students in the use and analysis of Galileo testing data. Teachers at one school initiated a student led review of assessment results. Students review each item and are able to deduce which standard each item tests, using both their own results and the student language standards. Teachers were trained to identify test item "distracters" to understand why students may have incorrectly answered a question. As one teacher explained, "Students discuss what they get and what they don't. It really helps them understand why I teach the stuff we have to focus on." One school has started using "daily math journals" that incorporate objectives and assignments that are mutually agreed upon by the student and teacher. Finally, many teachers have begun to use Galileo as part of their own assessment regimen by counting results as part of their student grading routine.

Galileo formative testing also is used by FPS to target intervention groups. Middle school math teachers used this formative assessment function to create a baseline math assessment. Results were used to homogeneously group students by math ability. The students were grouped in this manner not in their math courses, but rather in a "math seminar." Every 7th grade teacher teaches one of the math seminars that take place daily for 20 min. The 7th grade math teacher works with other teachers to assist them in developing targeted lessons based on the student needs in each group.

Now in their third year, Franklin has continued to revisit their curricular work to further "unwrap" the standards. Galileo results have driven this change. With review

of student learning by standard, teachers have made continual changes to both scope-and-sequence and power-standards. Both teachers and administrators have viewed the curriculum as malleable and as a work in progress. As the math curriculum specialist explained of Galileo, "It's a good judge of curriculum by classroom, and as a comparison of classroom curriculum to state curriculum. It's a driving force."

5 Evaluation of fit

Franklin, Ryder, and Woodbury schools invested in formative assessment systems in response to new accountability demands and to ultimately advance student achievement. In each district there was a belief that a well-designed formative assessment system would provide educators with information to advance their practice. Specifically, there was an underlying assumption that educators, especially teachers, needed more information about student learning to modify their work. In the previous section, we described each assessment system in regard to its development, purpose, technical features, and enactment in each district. In this section, we discuss the successes and shortcomings of each district's application of their chosen assessment system.

5.1 Ryder's fit

While the assessments developed in Ryder reflect a genuine effort to create a formative assessment, the current strategy of creating tests that include content already taught and content yet to be taught reduces the effectiveness of each assessment. Specifically, there are too few items per standard on each test to accurately measure students' proficiency with respect to past, current, or future curriculum. Several other factors compound this shortcoming and render the system even less valid with respect to diagnosing student need. The benchmark assessments used in Ryder may differ in difficulty and so a higher score on an easier test may not reflect better performance relative to a more difficult test. When tests of different difficulty are administered at different points in time, there is no way to know if learning has occurred over that time period. Similarly, if the same items are used across different time periods, there is no way to separate the degree to which student learning or a practice effect (due to item familiarity) explains changes in test performance. Finally, because the items from these assessments are not calibrated onto a common scale, there is no way to measure growth within or across a school year.

Evaluating the validity of the Ryder assessments requires considering their intended purpose. Several purposes were stated by district staff, including providing practice to students for the state assessment, holding teachers accountable for their instruction, monitoring students' progress on the benchmarks, facilitating constructive discussions among teachers, and predicting students' performance on the state assessment. With respect to providing practice to students and facilitating discussions among teachers, it is likely the assessments are fulfilling those goals. The students are exposed to state assessment questions four times before they take the State Assessment in the spring, and teachers meet after each assessment to

discuss the results. Discussions among administrators and teachers may also be aided by reviewing test results and so these assessments may also have utility as part of a teacher accountability mechanism, since all teachers within the district who teach the same subject in the same grade use a common assessment.

With respect to the other two goals—monitoring students progress and predicting state assessment performance—there is (currently) little evidence to support its use. To measure educational gain over time requires a common scale for the assessments. Each pretest and QA is uniquely developed; however, they are developed to be strictly parallel. The first item on the first QA is considered equivalent to the first item on all subsequent QA, the second item on the first QA is assumed to be equivalent to the second item on the subsequent QA, and so forth. This equivalence of items across assessments is attempted by either using the same item across assessments, or using a variation of the original item on subsequent assessments. In a sense, the attempt is to "clone" an original item so that it looks different on subsequent assessments, but has roughly the same difficulty. Using this strategy, the primary report reviewed by teachers is a graph of the proportion of students who answered each item correctly across assessments.

When an item is changed from one assessment to another to disguise its content, it cannot be assumed it retains its difficulty and so comparing the proportions of students who correctly answered an item over administrations is inappropriate. This problem is particularly prevalent in reading assessments where different passages are used. However, even in math, changes that appear to be superficial are known to have the potential to dramatically alter the difficulty of an item (Carlson and Ostrosky 1992). If the same item is used across administrations, it is difficult to disentangle any practice effect students may have from previously seeing the item, from real learning. Thus, the Ryder assessments are limited in the information they provide for measuring students' learning over time. Scores from the Ryder assessment are not on a standardized score scale, which makes comparisons across administrations difficult. The value of these assessments seems to be in providing information on benchmarks recently or soon to be taught. Nevertheless, it is somewhat inefficient, and possibly unethical, to repeatedly test students on benchmarks that have not yet been taught.

Finally, this assessment is not suited as a predictive model as intended by the district. For the Ryder assessment to predict students' state assessment performance, a predictive validity study is needed. The study would need to gather pretest, QA, and state test scores for the same group of students to derive prediction equations. In the end, Ryder had good intentions, but failed to meet all the goals they wanted from a FAS.

5.2 Woodbury's fit

Like all formative assessments, evidence based on test content is critical for the MAP to support diagnostic or formative inferences. The processes used to customize MAP test content provides evidence of a sound process, but the congruence of MAP content to state and local curricula should be verified for each state or district. Staff from Woodbury did not recall any content validity or alignment documentation specific to the state. Our evaluation of the description of validity presented in the

NWEA *Technical Manual* and the data presented led us to conclude MAP scores are generally valid for their intended purposes, particularly for measuring student growth over time. However, more information is needed to certify their appropriateness for making curricular decisions at the local level.

The MAP assessments from NWEA also tailor the test to match curricular needs, but the level at which the tailoring is done is more general. By maintaining a general focus, NWEA is able to link local MAP assessments to national norms and to a uniform score scale that is appropriate for measuring change across time. As a result, the MAP assessments can be used as an effective tool for the purpose of tracking student growth over time. Consequently, NWEA-MAP fit the stated district-level needs of Woodbury. That is, NWEA-MAP was a valid measure for student growth that was the state purpose of the Woodbury district. However, as our findings suggested, both the district purpose and assessment validity did not provide formative features for teachers. That is, teachers were not provided student diagnostic data that could have informed their practice. Moreover, by attempting to use this summative growth-data to alter instructional programming, Woodbury engaged in solutions based on student-level data that did not exist. For example, teachers were making instructional decisions such as homogenous grouping using data that did not generate such diagnostic student data. As a result, there was evidence of fit between central office personnel's desire for student growth data and the characteristics of the MAP system. However, there was no fit for teachers in the district seeking student achievement data to inform their practice.

5.3 Franklin's fit

Our research indicated that the Galileo formative assessment system was aligned with the Franklin district's intended use. We base this conclusion on: (1) the district working collaboratively with the vendor to select the items that best match the specific standards to be tested, (2) alignment of the assessments with the district curriculum (which is aligned with the state curricular frameworks), and (3) students are tested on standards that were most recently taught. With respect to the first point, this review feature is particularly impressive because teachers and curriculum leaders in the district can reject specific items and request that they be replaced for reasons of poor alignment with instruction or any other perceived problem. This review process helps closely align the benchmark assessments with instruction within the district, and allows teachers to discover how well students perform with respect to the specific knowledge and skills they are teaching. Consequently, the Galileo benchmark assessments are likely to have adequate content validity for student-level diagnostic information, although studies of the degree to which the assessments actually do represent the intended curriculum should be conducted. In the end, we found a strong fit between what Franklin wanted out of an assessment and the Galileo assessment system.

5.4 Summary of fit

Each of the assessment systems used in these three districts have strengths and limitations. We judge the MAP assessments available through NWEA as having

high technical quality for measuring student growth over time and comparing students to national norms. The Galileo benchmark assessments appear to be useful for measuring content specified by a school district and to provide a summary of how well students master content that was recently taught. In Ryder there was a disconnection between the assessment instrument and the intended purposes. While the district clearly had a formative intent, classroom educators were unable to use the Quarterly Assessment to inform their teaching.

Table 2 below summarizes our findings related to fit including valid system characteristics and the intended and actual uses by district educators.

6 Implications for schools seeking formative assessment systems

The current accountability pressures have been manifested in schools as assessments. As a result, the current educational reality for educational accountability stipulates that the assessment "tail [is] definitely…wagging the curriculum/instruction canine" (Popham 2004, p. 420). This study indicates the importance for schools considering the implementation of FASs to consider the issue of fit. The investigation of formative assessment tools and district intended and actual use

Table 2 Three formative assessment systems within the fit framework

FAS	Valid characteristics	Stated intent	Actual use	Fit
District-Made (Ryder)	Measuring pre-post instructional differences	• Assessment of taught curriculum • Competencies of untaught curriculum • State assessment prediction	• Limited re-teaching • Curricular monitoring	No
NWEA's MAP (Woodbury)	Evaluating growth General content inferences	• Baseline student diagnostic data • Student-growth data • Assessment of taught curriculum	• Student long-term grouping • Curricular monitoring • Student growth monitoring	Partial
ATI's Galileo (Franklin)	Measuring what was recently taught Specific content inferences	• Assessment of taught curriculum • Generate conversations and actions around student learning and instruction • Timely data • Data warehouse	• Student short-term grouping • Differentiated instruction • Re-teaching • Pedagogical collaborations • Curricular monitoring	Yes

must be conducted simultaneously. Examinations of the characteristics of assessment systems in isolation will not help consumers interpret the idealized functionality of the system's prowess through the unique context of each setting. The lesson here resides in the clear understanding that the utility of assessment tools lies in the pairing of the characteristics of the assessment and the intended and communicated purposes of the district. Validating the use of a test for a particular purpose requires gathering evidence that the interpretations made on the basis of test scores are valid. Consequently, the alignment of the assessment system and the intended use may yield great prognostic powers in the determination of its effective and meaningful use. Without specific attention to both elements that make up fit may result in implementation frustrations at best and at worst inappropriate misuses of assessment data. However, merely obtaining student achievement data that fits a district's needs is not the sin quo non of meaningful and effective data use. The transformation of data to new knowledge for teachers, and the translation of this knowledge into one's pedagogy, will require more that good fit. Studies have illustrated that under norms of collaboration, data can be used in a non-threatening and effective manner (Wayman and Stringfield 2006; Young 2006). Where assessments focused on instruction, student diagnostics, and teacher professional development, educators developed efficacy toward assessments (Popham 2008; Stecher 2002). As a result, the implementation of a formative assessment system must be accompanied with the development of a data-driven professional learning community. Teacher professional development has always been anchored in specificity, relevancy, and consistency of both pedagogy and content (c.f., Hawley and Valli 1999; Little 2002). However, understanding how students learn the content is of equal importance in the teacher professional development process (Sykes 1999). Using assessments that have the " right fit" may help focus professional development efforts in order to create a data-driven professional learning community.

While not the focus of this study, our research indicated that, in conjunction with our concept of fit, a number of organizational elements might have impacted the meaningful and appropriate use of formative assessments in schools. For example, the development of district and school level data teams are useful to: Complete curricular and assessment inventories, identify professional development needs, and incorporate rituals of use that are monitored and supported. In the end, fit must be accompanied by coherence in the organization and the development of educator's capacity.

7 Conclusion

Schools should be held accountable to individual student learning over a period of time (Elmore 2004). However, the high demand to use data coupled with the inadequate training and pervasive fear of results places emphasis on the assessments themselves and not on teaching and learning (Earl and Katz 2002). Consequently, the thoughtful, effective use of student assessment data has yet to be realized. The relentless pressures for schools to show student growth *within* the school year combined with the burgeoning formative assessment system market make this topic timely and ripe for investigations such as the present study.

The formative assessment systems used in these three school districts serve several purposes including evaluating and informing curricular change, diagnosing students' strengths and weaknesses, and measuring growth. Interpretations of the results of these systems are made at several different levels including the student, classroom, school, and district. Thus, evaluating the decisions and actions that are made on the basis of the results from these assessments involves evaluating the evidence associated with each assessment that would support each specific activity. In the end, both use (intended and actual) and system characteristics matter.

Each formative assessment system has a utility. In addition to the required state summative assessments, there is often a place for a nationally norm-referenced assessment, a student growth model assessment, and or a local formative assessment system. However, districts must be strategic in the implementation of new assessments. There is a clear danger in over testing students and overwhelming teachers with information. What is clear is that fit matters. That is, districts must clearly articulate and understand their data needs and find a formative assessment system that addresses their needs in valid ways. Consequently, districts must first ask: Why do we want a formative assessment system? What do we want from the system? Who needs this information (e.g., teachers, district and school administrators, parents, students)? How will the results drive curricular, developmental, and instructional modifications? Simply put, the intent of using data to inform and mediate pedagogical practice and administrative decision-making in schools is laudable, however without understanding fit in regard to formative assessment system capabilities such espoused ends are untenable. As schools search for and fend off formative assessment systems, they would be well advised to arm themselves with the importance of fit.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.

Baker, E. L. (2007). The end(s) of testing. *Educational Researcher, 36*(6), 309–317.

Bergan, J. R., Bergan, K., Guerrar Burnham, C., Cunningham, S., Feld, J., Nochumson, K., & Smith, K.A. (2006). *Building reliable and valid benchmark assessments*. Tucson, AZ: Assessment Technology Incorporated.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–74.

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

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

Bogdan, R. C., & Biklen, S. K. (1992). *Qualitative research for education: An introduction to theory and methods*. Boston: Allyn & Bacon.

Brunner, C., Fasca, C., Heinze, J., Honey, M., Light, D., Mandinach, E., et al. (2005). Linking data and learning: the grow network study. *Journal for Students Placed at Risk, 10*(3), 241–267.

Carlson, J. L., & Ostrosky, A. L. (1992). Item sequence and student performance on multiple-choice exams: Further evidence. *The Journal of Economic Education, 23*(3), 232–235.

Coburn, C., & Talbert, J. (2006). Conceptions of evidence use in school districts: mapping the terrain. *American Journal of Education, 112*(4), 469–495.

Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks: Sage.

Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record, 106* (6), 1047–1085.

Datnow, A., Park, V., & Wohlstetter, P. (2007). *Achievig with data: How high-performing school systems use data to improve instruction for elementary students*. Los Angeles: Center for Educational Governance at University of Southern California.

Earl, L., & Katz, S. (2002). Leading schools in a data-rich world. In K. Liethwood & P. Hallinger (Eds.), *Second international handbook of educational leadership and administration, Part Two* (pp. 1003–1023). Dordrecht: Kluwer.

Elmore, R. (2004). The problem of stakes in performance-based accountability systems. In S. H. Fuhrman & R. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 274–296). New York: Teachers College Press.

Gitomer, D., & Duschl, R. (2007). Establishing multilevel coherence in assessment. In P. Moss (Ed.), *Evidence and decision making. 106th yearbook of the national society for the study of education, Part I* (pp. 288–320). Chicago: University of Chicago Press.

Halverson, R. (2003). Systems of practice: how leaders use artifacts to create professional community in schools. *Educational Policy Analysis Archives, 11*(37), 1–35.

Halverson, R., Grigg, J., Prichett, R., & Thomas, C. (2007a). The new instructional leadership: creating data-driven instructional systems in school. *Journal of School Leadership, 17*(2), 159–194.

Halverson, R., Prichett, R., & Watson, J. (2007b). *Formative feedback systems and the new instructional leadership*. Madison: Wisconsin Center for Education Research.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park: Sage.

Hawley, W., & Valli, L. (1999). The essentials of effective professional development: a new consensus. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 151–180). San Francisco: Jossey-Bass.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: American Council on Education/Praeger.

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

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

Leighton, J. P., & Gierl, M. J. (2007). Why cognitive diagnostic assessment? In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive assessment for education: Theory and applications* (pp. 3–18). New York: Cambridge University Press.

Little, J. W. (2002). Professional communication and collaboration. In W. Hawley (Ed.), *The keys to effective schools: Educational reform as continuous improvement* (pp. 43–55). Thousand Oaks: Corwin.

McDonnell, L. (2004). *Politics, persuasion, and educational testing*. Cambridge: Harvard University Press.

Merriam, S. B. (1988). *Case study research in education: A qualitative approach*. San Francisco: Jossey-Bass.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: Macmillan.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Beverly Hills: Sage.

Militello, M. (2004). At the cliff's edge: utilizing evidence of student achievement for instructional improvements. *Dissertation Abstracts International (AAT 3158978), 65*(12A), 4419.

Militello, M. (2005). *The construction and implementation of assessment accountability at the district level.* Paper presented at the University Council for Educational Administration, Nashville, TN, November.

Militello, M., & Heffernan, N. (2009). Which one is "just right"? What educators should know about formative assessment systems. *International Journal of Educational Leadership Preparation, 4*(3), 1–8.

Militello, M., & Sykes, G. (2006). *Why schools have trouble using data.* Paper presented at the National Council on Measurement in Education, San Francisco, CA, April.

Militello, M., Sireci, S., & Schweid, J. (2008). *Intent, purpose, and fit: An examination of formative assessment systems in school districts.* Paper presented at the American Educational Research Association, New York City, NY, March.

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

Northwest Evaluation Association. (2005). *Technical manual: For use with measures of academic progress and achievement level tests*. Lake Oswego: NWEA.

Ogawa, R. T., & Collom, E. (2000). Using performance indicators to hold schools accountable: implicit assumptions and inherent tensions. *Peabody Journal of Education, 75*(4), 200–215.

Ogawa, R. T., Sandholtz, J., Martinez-Flores, M., & Scribner, S. P. (2003). The substantive and symbolic consequences of a district's standards-based curriculum. *American Educational Research Journal, 40* (1), 147–176.

Pellegrino, J. W., Chudowsky, N., & Glaser, B. (2001). *Knowing what students know: The science and design of educational assessment*. Washington: National Academic Press.

Perie, M., Marion, S. F., & Gong, B. (2007). *Moving towards a comprehensive assessment system: A framework for considering interim assessments*. Dover: The National Center for the Improvement of Educational Assessment, Inc.

Petrides, L. A., & Guiney, S. Z. (2002). Knowledge management for school leaders: an ecological framework for thinking schools. *Teachers College Record, 104*(8), 1702–1717.

Popham, W. J. (2004). Curriculum, instruction, and assessment: amiable allies or phony friends? *Teacher College Record, 106*(3), 417–428.

Popham, W. J. (2008). *Transformative assessment*. Baltimore: ASCD.

Sharkley, N. S., & Murnane, R. (2006). Tough choices in designing a formative assessment system. *American Journal of Education, 112*, 572–588.

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

Stecher, B. (2002). Consequences of large scale, high stakes testing on school and classroom practice. In L. S. Hamilton, B. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 79–100). Santa Monica: RAND.

Stiggins, R. (2005). From formative assessment to assessment FOR learning: a path to success in standards-based schools. *Phi Delta Kappan, 87*(4), 324–328.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park: Sage.

Streifer, P. A., & Shumann, J. S. (2005). Using data mining to identify actionable information: breaking ground in data-driven decision making. *Journal of Education for Students Placed at Risk, 10*(3), 281–293.

Sykes, G. (1999). Teacher and student learning: strengthening their connection. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 151–179). San Francisco: Jossey-Bass.

U.S. Department of Education. (2003). Using data to influence classroom decisions [Electronic Version]. Retrieved July 28, 2007 from http://www.ed.gov/teachers/nclbguide/datadriven.pdf.

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

Wayman, J., & Stringfield, S. (2006). Data use for school improvement: school practices and research perspectives. *American Journal of Education, 112*(4), 463–468.

Weiss, C. H. (1995). The four "I's" of school reform: how interests, ideology, information, and institution affect teachers and principals. *Harvard Educational Review, 65*(4), 571–593.

Wiliam, D. (2006). Formative assessment: getting the focus right. *Educational Assessment, 11*(3&4), 283–289.

Young, V. M. (2006). Teachers' use of data: loose coupling, agenda setting, and team norms. *American Journal of Education, 112*(4), 521–548.