5-2012

In Pursuit of a Balanced System of Educational Assessment: An Evaluation of the Pre-Kindergarten Through 8th Grade Math Assessment System in One Massachusetts Regional School District

Rita Joyce Detweiler
University of Massachusetts Amherst

Recommended Citation

Detweiler, Rita Joyce, "In Pursuit of a Balanced System of Educational Assessment: An Evaluation of the Pre-Kindergarten Through 8th Grade Math Assessment System in One Massachusetts Regional School District" (2012). *Open Access Dissertations*. 542.
https://doi.org/10.7275/2n8w-5483 https://scholarworks.umass.edu/open_access_dissertations/542

IN PURSUIT OF A BALANCED SYSTEM OF EDUCATIONAL ASSESSMENT: AN EVALUATION OF THE PRE-KINDERGARTEN THROUGH 8TH GRADE MATH ASSESSMENT SYSTEM IN ONE MASSACHUSETTS REGIONAL SCHOOL DISTRICT

A Dissertation Presented

by

RITA J. DETWEILER

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

May 2012

Educational Policy, Research and Administration

IN PURSUIT OF A BALANCED SYSTEM OF EDUCATIONAL ASSESSMENT: AN EVALUATION OF THE PRE-KINDERGARTEN THROUGH 8TH GRADE MATH ASSESSMENT SYSTEM IN ONE MASSACHUSETTS REGIONAL SCHOOL DISTRICT

A Dissertation Presented

By

RITA J. DETWEILER

Approved as to style and content by:

Rebecca H. Woodland, Chair

Kathryn A. McDermott, Member

Melissa S. Woodard, Member

Christine B. McCormick, Dean
School of Education

DEDICATION

To my husband, Rob Detweiler.
This accomplishment is in your honor.

And to our children, Ayana, Jim, and Laurel Detweiler
and the wonderful new additions both now and in the future
Meagan Moos and Grace Detweiler

ACKNOWLEDGMENTS

I want to thank Dr. Rebecca Woodland, my adviser and dissertation chair, for your support throughout this process. There were definitely times I might have abandoned the whole endeavor if you had not encouraged me to continue. I enjoyed our many planning sessions and always walked away from them with new and creative ideas. I want to thank Dr. Kathryn McDermott and Dr. Melissa Woodard. You were excellent teachers during my course work with you. It has been my privilege to work with you as members of my committee. I thank all of the members of the administrative team and the staff in the district in which I conducted this study. Your willingness to reflect on your process is a model for other school districts who are seeking to change and improve.

Many thanks to my wonderful husband. You were a great sport to go along with this idea in the first place. You and I both know how generous and unselfish you were in supporting me along the way. Now it is your turn to take us off on the next leg of our journey together!

ABSTRACT

IN PURSUIT OF A BALANCED SYSTEM OF EDUCATIONAL ASSESSMENT:
AN EVALUATION OF THE
PRE-KINDERGARTEN THROUGH 8TH GRADE MATH ASSESSMENT SYSTEM
IN ONE MASSACHUSETTS REGIONAL SCHOOL DISTRICT

MAY 2012

RITA J. DETWEILER, B.A., EARLHAM COLLEGE

M.Ed., UNIVERSITY OF MASSACHUSETTS AMHERST

C.A.G.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Associate Professor Rebecca Woodland

School leaders in the United States live in an educational era characterized by a desire for and expectation that all students attain high levels of academic proficiency. There is an increased reliance on all types of educational assessment as a key component to help school leaders attain that goal. The purpose of this study is to understand how school administrators can foster a balanced system of assessment at the local level to genuinely harness the power of assessment to enhance student learning. The significance of the study rests in the fact that there is a general failure of states and school districts to conceive of educational assessment as a system that operates across levels of the educational system from the classroom on up to the district and state level. The findings of this study are intended to support the efforts of a group of administrators to develop a balanced system of math assessments in their school district.

TABLE OF CONTENTS

Page

CHAPTER

APPENDICES

LIST OF TABLES

TABLE OF FIGURES

CHAPTER 1

INTRODUCTION

Statement of the Problem

School leaders in the United States live in an educational era characterized by a desire for and expectation that all students attain high levels of academic proficiency. The use of educational assessment as a means to help school leaders reach that goal is a key feature in contemporary school reform efforts (Hamilton, Stecher, & Yuan, 2008; Pellegrino & Goldman, 2008; Ryan, 2002). The challenge for school leaders is to understand and employ educational assessment in ways that genuinely enhance student learning.

The educational assessment of students refers to the process of reasoning from evidence about student learning. This process involves developing measures that are "designed to observe students' behavior and produce data that can be used to draw reasonable inferences about what students know" (Pellegrino, Chudowsky, & Glaser 2001, p. 42). Many factors will have an effect on the design of specific assessment measures, including decisions about the nature of learning and what constitutes evidence.

Educational assessment is utilized for a variety of purposes. Assessment can provide evidence of student achievement at the end of a learning sequence to determine if a student has achieved a level of mastery (Black, 1993b; Harlen & James, 1997; Stiggins, 1995; Taras, 2005). When assessment is used for this purpose, it is typically referred to as the summative use of assessment. Measures, such as end-of-unit tests and large-scale standardized assessments, are commonly employed by educators to provide this type of evidence.

Assessment can provide evidence of student progress to inform the day-to-day decisions that shape the on-going teaching-learning experience (Black & Wiliam, 1998, 2009; Chappuis, 2009; Earl, 2003; McMillan, 2007). When assessment is used for this purpose, it is typically referred to as the formative use of assessment. Educators use measures, such as classroom observation, teacher-student conferences and student self-assessment, to yield this type of evidence (Andrade & Boulay, 2003; Fernandez & Fontana, 1996; Sadler, 1989; Stiggins, 2001).

Assessment can also provide evidence that enables school administrators and policy makers to make decisions about the quality and effectiveness of educational programs and personnel (Council of Chief State School Officers [CCSSO], 2002). This use of assessment is typically referred to as the evaluative use of assessment. The National Assessment of Educational Progress (NAEP) is one example of such a measure. NAEP assesses broad trends in achievement for students nationwide and provides an independent source of information about how students in participating states are performing relative to the nation as a whole.

The increased reliance on all types of educational assessment as a key component of school reform efforts has led to efforts to understand how assessment practices can genuinely enhance student learning. One area of study has focused on the effect of assessment measures that are intentionally organized into a balanced system of assessment (Chappuis, Commodore, & Stiggins, 2010; Pellegrino & Goldman, 2008; Rothman, 2010). A balanced system of educational assessment is considered to have a *comprehensive range* of assessments, implying that there is a full range of measures that are used for summative, formative, and evaluative purposes that are administered with different frequencies throughout the teaching-learning cycle (Chappuis et al., 2010; Perie, Marion, Gong, & Wurtzel, 2007). A balanced system has *coherence amongst components.* This implies that the educational assessments are aligned to other components of the system, including curriculum and instruction, and all components reference a core set of standards that reflect developmentally appropriate learning sequences (Pellegrino & Goldman, 2008; Rothman, 2010; Shepard, 2000). A balanced system has a *robust capacity for data management* that enables a variety of stakeholders, including educators, policy makers and parents, to access the results of assessment to inform key decisions. Students are also empowered to understand and use assessment results to support their own learning (Boudett, City, & Murnane, 2005; Boudett & Steele, 2007; Love, Stiles, Mundry, & DiRanna, 2008). The assessment system incorporates measures that are *high-quality and diverse* to ensure that all students, including those who have been identified with learning disabilities or are from cultural, linguistic, or racial minorities, can be accurately and fairly assessed (Huai, Braden, White, & Elliott, 2006). A well-balanced system also places a *minimum burden* on students and staff to develop, obtain, analyze, interpret, and use assessment information (Boudett & Steele, 2007).

The efforts to understand the effect of a balanced system of assessment on student learning is hampered by the reality that a balanced system of assessment is not common practice. Pellegrino and Goldman (2008) note that

> Across the country there has been a general failure of states and school districts to realize that assessment has a very powerful and beneficial role to play in the instructional process, but only when it is conceived as a system that operates across levels of the educational system from the classroom on up to the district and state level with appropriate information flow in both directions. Furthermore, there is a general failure to realize that such a system requires multiple components, each of which is designed to assist the key actors at each level of the

> system by providing appropriate assessment tools that yield actionable information at that level. (p. 38)

Efforts to foster the development of balanced systems of assessment are a needed first step in the process of understanding the effect of these systems on student learning.

School administrators play a pivotal role in the development of balanced assessment systems in their school districts. Chappuis et al. (2010) assert that the “locus of control for the achievement of assessment balance and control is the local school district, as this is the only level of the educational system at which assessment can serve valuable purposes at annual, interim/benchmark, and classroom levels” (p. 25). The expectation that school administrators will be actively involved in the development of assessment systems is also reflected in the standards of performance for education leaders (CCSSO, 2008). These standards, initially articulated by the Interstate School Leadership Licensure Consortium (ISLLC) in 1996 and updated in 2008 as the *Educational Leadership Policy Standards: ISLLC 2008,* are a point of reference for state policy makers as they set guidelines for the preparation, licensure, evaluation, and professional development of school administrators (CCSSO, 2008). Embedded throughout these standards is the expectation that school administrators will develop assessment and accountability systems to identify goals, assess effectiveness, and monitor student progress.

Classroom teachers are critical actors in implementing and interpreting the assessment measures. In 1987 the American Federation of Teachers (AFT), the National Council on Measurement in Education (NCME), and the National Education Association (NEA) jointly undertook the task of articulating standards of assessment competency for teachers. The impetus for their work was the acknowledgement that “good teaching cannot exist without good student assessment" (AFT, NCME, & NEA, 1990, p. 1). Their work culminated in *The Standards for Teacher Competency in the Educational Assessment of Students*. The standards are intended to be guideposts for pre-service training programs and in-service professional development to ensure assessment literacy. Although dated, these standards are still the primary point of reference that articulate the skills that teachers need to ethically and appropriately develop, administer, score, interpret, and use assessment measures.

In summary, the use of educational assessment is a key feature of school reform efforts; however, the challenge for school leaders is to use assessment in ways that genuinely enhance student learning. Currently there is a general failure on the part of state and local school districts to realize how critical it is to conceive of assessment as a system that operates across all levels of the educational system (Pellegrino & Goldman, 2008). School administrators are in a unique position to foster a more balanced system of assessment and to ensure that their staff is assessment literate and can appropriately implement the system they develop.

Context of This Study

As school administrators undertake the task of developing a balanced system of assessment and increasing assessment literacy within their school districts, they need to analyze their current status. In short, they need to engage in an evaluation process. Evaluation, in general, is a demonstrated method for analysis and a means of building capacity (Smith & Freeman, 2002). Patton (2008) draws a fundamental distinction between evaluation and research.

> Basic scientific research is undertaken to discover new knowledge, test theories, establish truth, and generalize across time and space. Program evaluation is undertaken to inform decisions, clarify options, identify improvements and provide information about programs and policies within contextual boundaries of time, place, value and politics. Research aims to produce knowledge and truth. Useful evaluation supports action. (p. 40)

A utilization-focused evaluation (UFE) is a type of evaluation that can be highly tailored to local conditions. Patton (2008), considered the father of UFE, notes that a UFE is specifically "done for and with specific intended primary users for specific, intended uses" (p. 37). Primary users, also referred to as stakeholders, work collaboratively with the investigator to define the questions that will guide the evaluation.

A UFE is also distinct from other types of evaluation by the extent to which its value is gauged by its utility, implying that the findings are intended to lead to real change. These characteristics of a UFE make it a suitable choice for school administrators as they analyze the current status of their assessment system and chart a course of action that leads to improvements.

In this study I undertake a UFE and work with a group of district and building level administrators to analyze their current math assessment system and the assessment literacy of staff. The catalyst for this work for these administrators was their analysis of their 2010 Massachusetts Comprehensive Assessment System (MCAS) results in Mathematics from which they concluded that students in their elementary schools were not performing at target levels. As part of their efforts to increase students' level of performance, they concluded that they needed to review their current math program. Reviewing their math assessment practices was going to be a key component of their overall analysis.

From 2008-2010 they had committed most of their efforts towards improving their literacy program and consequently had a good perspective on the resources that they would have to commit to the review of their math program. When I approached the superintendent of the school district with my proposal to help them analyze their assessment system, the superintendent was very receptive. She presented my proposal to the rest of the administrative team, including the Assistant to the Superintendent, the Curriculum Director for Elementary Education, the Building Principals from each of the elementary schools and the combined middle/high school, the Special Education Director, and the Technology Specialist. They were unanimous in their willingness to incorporate my work in their district. Although they did not have any previous experience with a UFE, they all committed to working within this framework as it appeared to be appropriate for their needs.

Over the course of several months and several planning meetings, I developed the plan for this UFE through extensive collaboration with this administrative group. The educational assessment of students was the broad focus for the evaluation as it was a match between my area of expertise and a general need within the district. The administrative team narrowed the focus of this evaluation to just their math assessment system given their current priorities. They also narrowed the focus to just grades pre-kindergarten through 8th grade with a particular focus on the transition between 6th and 7th grade. The rationale for this decision was that the district experiences an influx of in-state and out-of-state students to such an extent that approximately 20%, of the incoming 7th grade students are new to the district. The superintendent noted that precious instructional time is lost because the 7th grade faculty has to assess the skill levels of these new students in order to properly place them. These administrators were also very interested in understanding the current level of assessment literacy. They were well aware that the success of a program often relies on the expertise of the staff that are expected to implement it. In addition, a change in practice often entails professional development in targeted areas to develop the needed expertise.

Evaluation Research Questions

This administrative team and I collaboratively developed the evaluation research questions over the course of several planning sessions. The questions aim to analyze the current status of their system of math assessments and the levels of assessment literacy of their staff. The questions are as follows:

> Question #1: To what extent do we currently have a balanced system of math assessments in grades pre-kindergarten through 8th grade?

This question reflects the administrators' interest in analyzing their current math assessment system in relation to the characteristics of a balanced system of assessments. This will entail analyzing their current system along the dimensions of a comprehensive range of assessments, coherence amongst components, a capacity for data management, the diversity and caliber of the assessments, and the overall organization of the system to determine if it places a minimum burden on staff and students to develop, obtain, analyze, interpret and use assessment information.

> Question #2: To what extent are our 6th and 7th grade teachers using math assessment to facilitate the transition of continuing and in-coming students into 7th grade?

This question reflects the unique challenges of integrating a significant number of new students into their district at the 7th grade level. This influx of students creates an inordinate need for assessment information in order to properly place students in appropriate academic programs.

> Question #3: What is the level of competency of our staff relative to established standards of competency for the educational assessment of students?

The aim of this question is to obtain information about the current levels of assessment literacy of their classroom teachers. These administrators acknowledged that the successful implementation of their current and any future assessment system is highly dependent on the expertise of their staff.

In this study I work with this administrative team and district staff to gather data related to their three primary questions, to analyze and report findings to the administrative team, and to promote the use of the findings to make informed decisions going forward. The administrative team has identified several ways they can utilize the findings in their strategic planning for the 2011-2012 school year: 1) to inform decisions about the allocation of district resources, 2) to help set priorities for their district professional development program, and 3) to improve the transition of all in-coming students into their middle school.

In the remainder of this dissertation I review the research literature in relevant areas including a review of the dimensions of a balanced system of assessment (including an in-depth look at the summative, formative and evaluative use of assessment) and a review of the current standards of competence for teachers in the educational assessment of students. I explain in depth the methodology of my study. Through a thorough analysis of results, I address each research question and draw conclusions that can inform the district administrative team as they chart a course of action. I end with implications for practice, policy, and future research.

CHAPTER 2

LITERATURE REVIEW

Introduction

School administrators have the challenge of designing systems of educational assessment that genuinely enhance student learning. A balanced system of assessments needs to provide policy makers, educators, parents, and students with the relevant data they each need to inform key decisions. To successfully implement a system of assessment, educators need to be assessment literate, implying that they can have the necessary skills for the ethical development, administration, scoring, and interpretation of assessment measures. The overarching goal for implementing a system of assessment is to ensure that assessment is used in such a way that it truly harnesses the power of assessment to enhance learning and enables all students to achieve high levels of academic proficiency.

Defining Educational Assessment

Learning is a complex internal mental process that cannot be directly perceived. The term "educational assessment" implies the use of less direct measures in educational contexts in an attempt to capture aspects of the very complex act of learning. Popham (2006) defines educational assessment as the "process by which educators use students' responses to specially created or naturally occurring stimuli in order to make inferences about students' knowledge, skills, or affective status" (p. 4). Pellegrino et al. (2001) define assessment as a process "designed to observe students' behavior and produce data that can be used to draw reasonable inferences about what students know" (p. 42).

Other researchers have adopted different definitions of educational assessment that highlight their different perspectives. McDonald and Boud (2003) shift the emphasis away from educators as the primary consumers of information and highlight the role of the student in the process of assessment. They define assessment as the process by which students identify “standards and/or criteria to apply to their work and making judgments about the extent to which they meet these criteria and standards” (p. 221). Hattie and Timperly (2007) underscore that educational assessment is not a stagnant linear process that ends with gathering information to make inferences. On the contrary, educational assessment entails the active use of the information as feedback into the learning process. Feedback is defined as “information provided by an agent regarding aspects of one’s performance or understanding” (p. 81).

These definitions reflect the shifting nature of our understanding of educational assessment. Pellegrino et al. (2001) presents a conceptual framework that characterizes all assessments. They postulate that every educational assessment is based on a set of philosophical assumptions that influence all aspects of the design of the assessment and they identify three core elements. There is an element of *cognition*, implying that every assessment is grounded in a theory of how people learn and how knowledge and understanding progress over time. There is an element of *observation*, which refers to the assumptions about which kinds of observations are most likely to result in students manifesting important knowledge or skills. The third element is *interpretation*. This refers to the “assumptions about how best to interpret the evidence from the observations to make meaningful inferences about what students know and can do” (p. 20). The importance of this framework is that it illustrates how a shift in any one of the elements, such as a new theory of learning or advances in how we observe or interpret learning, can have a profound impact on the design of assessment.

Shepard (2000) postulates that a shift in learning theory is related to the development and design of some of the new assessment practices that emerged in the 20th century. She contends that the learning theory that dominated most of the 20th century was rooted in behaviorist traditions. In broad terms, these learning theories conceive of the child as a *tabula rasa* whose cognitive development is dependent on externally manipulated processes. Motivation is primarily supported by external agents. Within this paradigm, the primary function of educational assessment was to document learning outcomes at the end of a learning sequence.

Alternative learning theories emerged in contrast to the behaviorist traditions. These theories had their roots in the theoretical work of Jean Piaget and Lev Vygotsky. Shepard (2000) refers to these new theories as social-constructivist, borrowing from cognitive, constructivist, and social-cultural traditions and summarizes the pertinent assumptions of this theoretical perspective in regard to cognitive development:

- Intellectual abilities are socially and culturally developed
- Learners construct knowledge and understandings within a social context
- New learning is shaped by prior knowledge and cultural perspectives
- Intelligent thought involves “metacognition” or self-monitoring of learning and thinking
- Deep understanding is principled and supports transfer
- Cognitive performance depends on dispositions and personal identity. (p. 8)

This shift in learning theory, from a behaviorist to a social-constructivist tradition, supported new forms of assessment. Within this paradigm, assessment is part of the on-going dialogue between the teacher and the learner. The social interchange helps to shape the learning process.

In summary, our understanding and ability to define what it means to engage in the educational assessment of students is constantly evolving. The disparity between the definitions from a range of different researchers reflects the shifting state of current practice. Understanding the differences in the underlying learning theory and/or the techniques used to observe, interpret, or apply the information gleaned from assessments can be helpful in explaining some of the differences in how assessment is conceived. Although these differences in our conceptual understanding complicate the process of incorporating sound assessment practices, nonetheless educators are still charged with the task of using existing assessment measures in ways that genuinely enhance learning.

A Balanced System of Assessment

The increased reliance on educational assessment as a key component of school reform efforts has highlighted the need to rely on all types of educational assessment. One area of study has focused on the development of balanced system of assessment. (Chappuis et al., 2010; Pellegrino & Goldman, 2008; Rothman, 2010). See Figure 1 for the conceptual design of a balanced system of assessments. This design is based on a thorough review of the literature and is original to this dissertation. At the core of the design are the assessments that are used for formative, summative, or evaluative purposes. The location in the triangle of each assessment category reflects how frequently the assessment is administered with formative assessments administered most frequently and evaluative assessments least frequently. The defining characteristics of a balanced assessment system are displayed around the sides and each is of equal importance.

A balanced system of educational assessment is considered to have a *comprehensive range* of assessment, implying that that there is a full range of measures that are used for formative, summative, and evaluative purposes that are administered with different frequencies throughout the teaching-learning cycle (Chappuis et al., 2010; Perie et al., 2007). A balanced system has *coherence amongst components* implying that the curriculum, instructional strategies and assessment practices all reference a common core set of standards (Pellegrino & Goldman, 2008; Rothman, 2010; Shepard, 2000). There is a *robust capacity for data management* that enables a variety of stakeholders to access the data to inform key decisions.

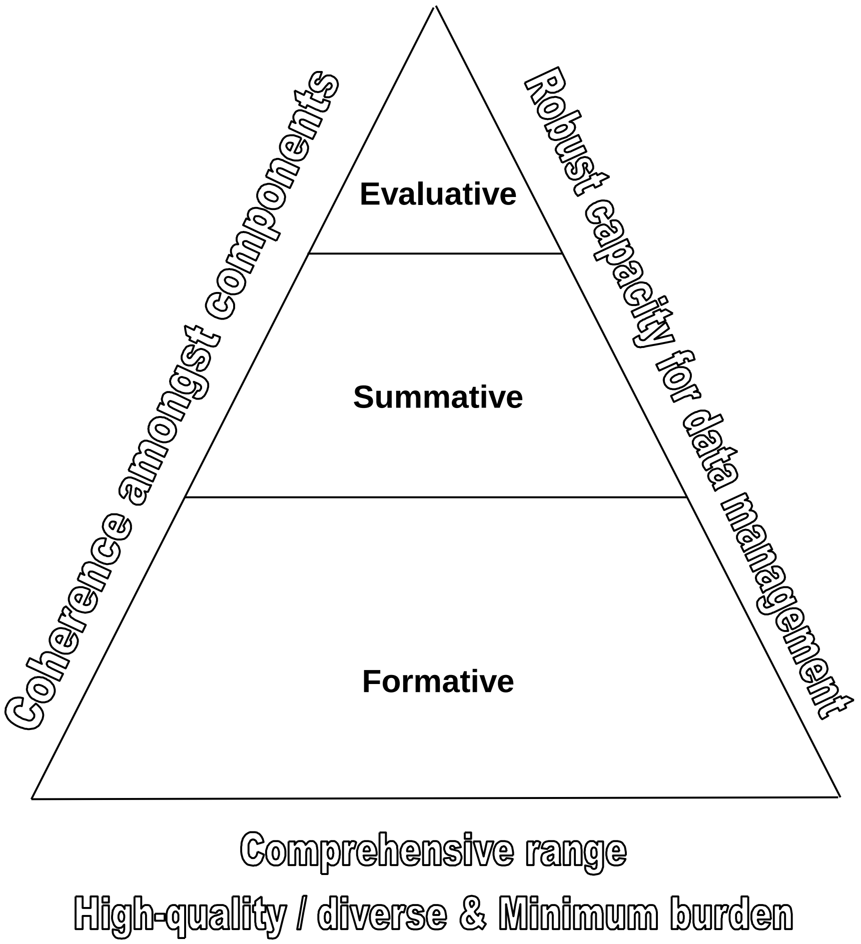

Coherence amongst components
Evaluative
Summative
Formative
Robust capacity for data management
Comprehensive range
High-quality / diverse & Minimum burden

Figure 1. A balanced system of assessment

Students are viewed as consumers of assessment data and empowered to understand and use data to support their own learning (Boudett et al., 2005; Boudett & Steele, 2007; Love et al., 2008). The system incorporates measures that are *high-quality and diverse* to ensure that all students, including those who have been identified with learning disabilities or are from cultural, linguistic, or racial minorities, can be accurately and fairly assessed (Huai et al., 2006). A well-designed system places a *minimum burden* on students and staff to obtain, analyze and interpret the assessment information (Boudett & Steele, 2007).

A Comprehensive Range

A comprehensive range of assessments implies that there are assessments that are used for a wide variety of purposes, including formative, summative, and evaluative. The formative use of assessment typically refers to the use of assessment measures to provide continuous feedback to teachers and students during instruction with the goal of modifying instruction and influencing student involvement in the learning process (Black & Wiliam, 1998, 2009; Chappuis, 2009; Earl, 2003; McMillan, 2007). The summative use of assessment typically refers to assessment measures to document achievement at the end of instruction with the goal of providing information in regard to level of mastery (Black, 1993b; Harlen & James, 1997; Stiggins, 1995; Taras, 2005). The evaluative use of assessment typically refers to the use of assessment measures to evaluate and make decisions about the quality and effectiveness of educational programs and personnel (CCSSO, 2002).

In addition to ensuring that there is a full complement of types of assessment, it is equally important to consider the quantity and frequency of administration of each type of assessment. In a well-balanced system, most of the assessments are used for formative purposes, and they are administered frequently. Fewer assessments are used for summative purposes and typically are administered less frequently. Assessments used for evaluative purposes are fewer still and typically require comparing data gathered over a period of months and years. See Table 1 for a summary of the purpose and frequency of assessment. A new hybrid measure, referred to as interim benchmark assessments, will be reviewed but is not incorporated into this model due to its uncertain value (Perie, 2007).

Table 1
Frequency and Purpose of Assessment

Formative use	**To provide continuous feedback to teachers and students during instruction with the goal of modifying instruction**		
	Examples	Intended use	Used by
Long-cycle End of year	Analysis of student portfolios from year to year	To assess student's rate of progress over the year	Teaching staff, parents, and students
Mid-Cycle 3-4 times/year	Benchmark using Dynamic Indicators of Early Learning Literacy Skills (DIBELS)	To monitor progress and identify students in need of remedial help	Teaching staff, parents, and students
Short-cycle Daily-monthly	Frequent conferencing. Exit activity at end of class	To provide continuous feedback to shape learning	Teaching staff and students
Summative use	**To document achievement at the end of instruction with the goal of providing information in regard to level of mastery**		
	Examples	Intended use	Used by
Long-cycle End of year	State-mandated assessments (MCAS)	School accountability Graduation requirements	Policy makers School Admin. Teaching staff Parents and students
Mid-cycle 3-4 times/year	Final exams	To demonstrate mastery	School Admin. Teaching staff Parents and students
Short-cycle Weekly-monthly	Tests and quizzes	To document learning at the end of a unit of study	Teaching staff Parents and students
Evaluative use	**To evaluate and make decisions about the quality and effectiveness of educational programs and personnel**		
	Examples	Intended use	Used by
Long-cycle over one to multiple years	MCAS NAEP	Evaluation of programs and/or personnel	Policy makers School Admin.

The Formative Use of Assessment

When the primary purpose of an assessment is to provide continuous feedback during the teaching-learning cycle to modify instruction, it is typically referred to as the formative use of assessment (Black & Wiliam, 1998, 2009; Chappuis, 2009; Earl, 2003; McMillan, 2007). The formative use of assessment is a recent phenomenon, only gaining in popularity over the last two decades (Black & Wiliam, 1998; Shepard, 2008). Given its potential to positively affect student learning and the hurdles inherent in incorporating these new practices into a comprehensive system of assessment, the research on formative assessment will be extensively reviewed.

Early Research

The first reference to formative assessment appears in the research literature when Scriven (1967) utilized the term "formative evaluation" to refer to the practice of using evaluation to develop or improve an educational process. Bloom, Hastings, and Madaus (1971) adopted this term to refer to the on-going diagnostic tests that were part of their mastery learning model. They defined formative assessment as "the systematic evaluation in the process of curriculum construction, teaching and learning for the purpose of improving any of these three processes" (p. 117).

In these early years of conceptualization, Sadler (1983) contributed to our understanding of formative assessment through his work with college-aged students in Australia. Learning was conceptualized as a growth curve that compared the student's actual performance to the desired goals and conceived of the gap between the two as a shifting measure of competence. Echoing the central tenet of formative assessment, Sadler argued that “good evaluation is not adjunct to good teaching: it *is* good teaching” (p. 63).

Other pioneering researchers include Fuchs and Fuchs (1986) who introduced and researched the effect of curriculum based measurement (CBM). In their initial study into the effect of CBM on student achievement in reading, Fuchs and Fuchs operationalized CBM as “data collection that occurred at least twice each week, with decisions concerning the adequacy of programs formulated on an individual, not a group, basis” (p. 201). The data collection typically consisted of counting the number of words read correctly in one minute. These data would then be analyzed in relation to the typical rate of reading at different age levels and difficulty of text to generate a measure of academic progress over time. They concluded that students whose educational programs were systematically monitored and adjusted based on the use of CBM achieved 0.7 standard deviation units higher than students whose programs were not monitored and adjusted. When teachers were required to follow data-evaluation rules, such as gathering 7-10 data points to calculate the rate of mastery, accompanied with mandated changes in instructional methods if the rate was off target, student achievement increased 0.9 standard deviation units. These initial positive findings are often cited as evidence of the potential benefits and stoked interest in the formative use of assessment.

A few substantial reviews and meta-analyses were also published by the mid-1980s and provide insight into the early research on the formative use of assessment and the role of feedback on student performance. Natriello (1987) undertook a review of the research of assessment practices in schools and their effect on student outcomes. He based his analysis on 91 studies drawn from research in classroom and laboratory settings resulting in an eclectic mix involving both summative and formative practices. He concluded that the majority of the research into the effect of evaluation processes on student outcomes was irrelevant because key distinctions were conflated. Many studies did not control for the quality or quantity of feedback practices. He concluded that additional research needed to provide better descriptive accounts of how students were currently being evaluated under a variety of conditions.

Natriello's (1987) review of the research literature highlighted the important role of feedback in the assessment process. He defined feedback as "the communication of the results of the evaluation to relevant parties, including the students, parents, school officials, and potential employers" (p. 160). He noted that the form in which the feedback was communicated had an effect on student performance. Specifically, when feedback was restricted exclusively to the traditional use of grades, it was associated with a more pronounced stratification of students on measures of academic performance. The form in which feedback was communicated also had an affective value, impacting a student's sense of self-efficacy. He noted that further research was necessary in order to understand the relationship between feedback and self-efficacy.

The role of feedback in the teaching-learning cycle was also the focus of Kulik, Kulik, and Bangert-Drowns (1990). In their meta-analysis on the effectiveness of mastery learning programs, they reviewed 36 studies using Bloom's *Learning for Mastery*, which at that time was one of the few instructional programs that combined formative and summative assessment practices. They concluded that students in college, high school and upper elementary school who participated in mastery learning programs raised their final examination scores an average of 0.50 standard deviations. This marks an increase from the 50th to the 70th percentile as compared to students in control groups who were involved in traditional methods of instruction. Although both high and low-aptitude students improved their performance, the gains for low aptitude students were on the average of 0.61 standard deviations as compared to gains of 0.40 for high-aptitude students. Although the improved rates of achievement for low-aptitude students could not directly be attributable to the embedded elements of formative assessment, this research paved the way for other studies that could explore this connection.

Kluger and DeNisi's (1996) meta-analysis is one of the most thorough reviews on the effects of feedback interventions on performance. Although they concluded that feedback generally had a significant and positive impact on performance in the range of 0.50 standard deviations, however, about 40% of the studies reported negative effect sizes. To explain these contradictory findings, they identified moderators that could impact the effectiveness of the feedback and developed a theoretical model, Feedback Intervention Theory (FIT). FIT articulates five suppositions:

> (a) Behavior is regulated by comparisons of feedback to goals and standards
> (b) Goals and standards are organized hierarchically (task learning, task motivation and meta-cognitive tasks)
> (c) Attention is limited and therefore only feedback-standard gaps that receive attention actively participate in behavior regulation
> (d) Attention is normally directed to a moderate level of the hierarchy
> (e) Feedback interventions change the locus of attention and therefore affect behavior. (p. 259)

They postulated that when individuals are confronted with a gap between current performance and the target goal, they adopt one of the following four responses which function as moderators. An individual may 1) attempt to reach the standard or reference level. This is the typical response when the goal is clear and when individuals have a high commitment to achieving the goal and hold the belief that they can be successful. 2)

Abandon the standard completely. This often occurs when individuals hold the belief that success is unlikely. 3) Change the standard. Individuals choose this when they do not want to abandon the standard, but estimate the likelihood of success to be low. 4) Deny that the standard exists. Their research underscored that feedback has to be framed with reference to a goal or standard and needs to be understood from the perspective of the individual receiving the feedback.

Crooks' (1988) review of the research literature on classroom evaluation practices involved 241 studies, also comprising an eclectic mix of summative and formative practices. He noted that evaluation practices placed too much emphasis on summative assessments with the undesirable effects of reducing intrinsic motivation, increasing anxiety to levels that were debilitating for students, lowering self-efficacy in weaker students, and potentially increasing the likelihood of poor social relationships between students. He concluded that research indicated that feedback enhanced student achievement when it focused students' attention on their progress towards mastery. The effectiveness of feedback was enhanced when it was given soon after a task was completed and when it contained sufficient detail to enable the student to understand misconceptions and other shortcoming in performance.

There was some early interest in the use of student self-assessment, a novel way to use assessment for formative purposes. Through his on-going research with university students in Australia, Sadler (1989) noted that some students failed to develop competence even when instruction incorporated the formative use of assessments. He expanded the focus of his research onto student self-assessment. He noted that

> Providing guided but direct and authentic evaluative experience for students enables them to develop *their* evaluative knowledge, thereby bringing them

> within the guild of people who are able to determine quality using multiple criteria. It also enables transfer of some of the responsibility for making decisions from teacher to learner. In this way, students are gradually exposed to the full set of criteria and the rules for using them and so build up a body of evaluative knowledge. It also makes them aware of the difficulties that even teachers face of making such assessment, they become insiders rather than consumers. (p. 135)

His work contributed to our understanding of the value of incorporating students as active consumers of assessment information.

Fontana and Fernandez (1994) also focused their research on self-assessment and studied the three-way relationship between self-assessment, learning outcomes and perceived locus of control. They concluded that students who participated in self-assessment practices reported an increase in internal locus of control. The students attributed their success to internal factors, such as effort, compared to external factors, such as luck. Other factors, such as age, had a moderating effect on the shift in beliefs with nine-year-olds reporting a greater shift than eight-year olds. Fontana and Fernandez hypothesized that students have to reach a certain level of development before they can benefit from self-assessment.

Coming of Age

It is the seminal work of Black and Wiliam (1998), *Assessment and Classroom Learning,* which created a watershed moment that brought together the research community and introduced formative assessment to the public-at-large. They began their extensive review by noting several deficiencies in the research base. They noted that researchers were inconsistent in how they conceptualized and defined formative assessment. In an effort to pull together the various concepts, Black and Wiliam defined formative assessment as “encompassing all those activities undertaken by teachers and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged" (p. 7).

Black and Wiliam (1998) noted numerous shortcomings in the existing body of research on the effects of the formative use of assessment. They asserted that research had to be conducted in natural settings, such as the classroom environment, in order to truly measure the effects of the formative assessment; however, they could identify only a handful of studies which met this standard. They initially intended to conduct a meta-analysis of the research but could identify only 20 studies with sufficient rigor to be included in a typical meta-analysis. They therefore chose to conduct a less stringent review by including more studies on the grounds that they did not want to overlook "any important clues or pointers towards the difficult goal of reaching adequate complex and complete understanding of formative assessment" (p. 9).

In their analysis, they identified several key features of formative assessment. Formative assessment involves a sequence of actions; initially students have to perceive that there is a gap between the current level of performance and a desired goal and then they have to take some action to close the gap. Formative assessment involves feedback between the teacher and the student about progress relative to closing the gap. This exchange of feedback builds on the social nature of learning and empowers students as active participants in the learning process. Logically the feedback typically leads to a shift in instructional strategies to enhance the likelihood of closing the gap. Black and Wiliam (1998) assert that "it is not possible to introduce formative assessment without some radical change in classroom pedagogy because, of its nature, it is an essential component of the pedagogical process" (p. 10).

Black and Wiliam (1998) also threw the spotlight on self-assessment. Although they cited research that reported on the positive effect of self-assessment on achievement, they noted that “the focus on self-assessment by students is not common practice, even amongst those teachers who take assessment seriously” (p. 25). This gap between theory and practice typified the assessment practices at this time.

Despite some of the noted limitations in the research, Black and Wiliam’s (1998) analysis underscored the potential positive impact that the formative use of assessment had on student achievement. They reported positive effect sizes on student achievement in the range of .40 to .70 which is a much more profound and positive effect than typical educational interventions. Of equal importance, the positive effects manifested with a greater magnitude with low-achieving students compared to high-achieving students. Given the magnitude of the positive effect on student achievement, coupled with the potential to close the achievement gap between low and high-achieving students, interest in formative assessment practices increased in educational circles.

Contemporary Issues Regarding the Formative Use of Assessment

Theory and research into the effect of specific practices has continued in the 14 years since the publication of *Assessment and Classroom Learning* (Black & Wiliam, 1998). The challenge in deepening our understanding of formative use of assessment is that “formative assessment is not an instrument or an event, but a collection of practices with a common feature: they all lead to some action that improves learning” (Chappuis, 2009, p. 4). These collections of practices characterize a new culture of assessment and Black and Wiliam (2009) have identified five qualities that characterize a new culture. A culture of assessment that incorporates the formative use of assessment is one in which teachers, (1) clarify and share learning intentions and criteria for success, (2) engineer effective classroom discussions and other learning tasks that elicit evidence of student understanding, (3) provide feedback that moves learners forward, (4) activate students as instructional resources for one another and, (5) activate students as the owners of their own learning.

Our understanding of the role of feedback in the assessment process was the focus of a meta-analysis by Hattie and Timperley (2007). They concluded that when assessment provides feedback to teachers and students about the goals of instruction and progress towards those goals it has the potential to positively affect student achievement. The effect of feedback is mediated by differences in the capacity of learners to self-assess and their willingness to seek out feedback information. Effective learners engage in internal feedback while less effective learners are more dependent on external measures, such as teacher feedback. Feedback that is task-related and conveys specific information is more effective than feedback that is personal and interpreted as praise, characterized by comments such as “Good boy” or “Great effort.”

A new hybrid measure, referred to as interim benchmark assessment, is gaining in popularity (Goren, 2010; Pellegrino, & Goldman, 2008; Perie et al., 2007; Shepard, 2008). While the positive effect of formative assessment on student achievement is well-established in the research literature, there is “little research on what kinds of educational results can reasonably be expected from interim assessment and thus little evidence about the characteristics of an effective interim assessment system” (Perie et al., 2007, p. 7). There is also a growing concern about the scope that interim assessments will occupy in the overall balance of an assessment system. For school officials, incorporating interim assessments is a relatively straightforward process that often entails only the purchase of a commercially available product and a limited amount of supporting professional development. On the other hand, “because real formative assessment is so entwined with instruction and pedagogical process, much more sustained professional development and support are needed to help teachers make fundamental—and more effective—changes in their teaching practices” (Shepard, 2008, p. 298). Perie et al. (2007) echo these concerns and speculates that “one reason school districts are investing in interim assessment systems that they hope will serve instructional purposes, rather than promoting formative assessment, is that they lack the capacity to do formative assessment well at scale” (p. 17).

Despite these shortcomings and concerns, school districts are incorporating interim assessments into their comprehensive assessment systems at an increasing rate (Goren, 2010). School administrators need some criteria to aid in the development of these assessments at the local level or with the selection of commercially available products. Findings from recent research indicate that the unique “fit” between the needs of a school district and the formats of different interim assessment systems is an important consideration (Millitello, Schweid, & Sireci, 2010). Because one size does not fit all, school administrators need to identify their primary purpose for administering an interim assessment and then select a corresponding system accordingly.

In summary, research into the formative use of assessment is a relatively contemporary area of study. Findings from research indicate that the implementation of formative assessment practices have a significant and positive effect on student achievement with greater benefit to low-achieving students. The practical applications of formative assessment in the classroom are evolving. When the formative use of assessment is integrated into an overall system of assessments, it signals a fundamental shift in the culture of assessment.

The Summative Use of Assessment

When the primary purpose of assessment is to document individual student achievement at the end of the teaching-learning cycle, it is typically referred to as the summative use of assessment (Black, 1993b; Harlen & James, 1997; Stiggins, 1995; Taras, 2005). Quizzes, end-of-unit tests, and final exams are all measures that are typically used in a summative manner. Large-scale standardized tests, such as the Massachusetts Comprehensive Assessment System (MCAS), are also used for summative purposes. In some states the results from these large-scale tests are used to determine if a student has met requirements for promotion or graduation.

History of the Summative Use of Assessment

The summative use of assessment has dominated the assessment landscape throughout most of the 20th century and continues to be “one of the most sacred traditions in American education” (Olson, 1995, p. 24). When Stiggins and Bridgeford (1985) surveyed assessment practices of 288 teachers in 8 representative districts throughout the United States, they found that 50% of teachers reported using teacher-made, multiple-choice tests that were administered at the end of the teaching cycle. The reliance on these tests increased as grade level increased. Based on their research, Stiggins and Bridgeford concluded that the majority of teachers exclusively employed summative assessment practices, most often teacher-made objective tests.

The effect of the summative use of assessment on student learning has been researched (Brookhart, 1999; Dweck, 1986). There are research findings that support the conclusion that the use of quizzes and end-of-unit tests has some beneficial effect on student learning (Shepard et al., 2005). The benefit is related to three factors: (1) additional practice with curricular content when students engage in review in preparation for a test, (2) the test itself engages students in the mental processing of the curricular content, and (3) the test directs the attention of the students to the content and skills that are tested and that has positive implications for subsequent learning.

Research has also been conducted on the effects of large-scale standardized assessments on student learning, however, several factors complicate these studies. Hamilton et al. (2008) note that it is difficult to disentangle the effects of these assessments from the other initiatives that are happening concurrently. It is difficult to generalize the effect that these tests have on student achievement from state to state because of the variability in state accountability policies. At present, each state defines the parameters of proficiency for their assessments and consequently these measures can vary significantly from state to state. Another limitation is related to the range of skills assessed by these assessments. In one study that involved a review of large-scale assessments from nine states, approximately “30 percent of the mathematics assessment items in those states matched the content and cognitive demand of the mathematics standards’ expectations in fourth grade; and only 26 percent matched the standards at eighth grade” (Rothman, 2010, p. 3). At this level of correspondence, these assessments may not measure a sufficient amount of material from which to accurately infer achievement at the level of the individual student.

Despite these limitations, available research does yield some broad findings. Research has supported the conclusion that the format of these large-scale assessments affects classroom practices in positive ways (Hamilton et al., 2008). Educators report that they have adapted their classroom assessments to mirror the format of the state tests and "in states where tests include open or extended-response items and are focused on higher-level cognitive skills, teachers have reported positive changes to assessment practices and greater emphasis on the quality of their own classroom-level assessments" (Abrams, 2007, p. 85). Educators also report that large-scale assessments have resulted in

> adopting new programs that address the needs of low-performing students, aligning curriculum and assessment programs to state standards, increasing the use of data to improve decision making and providing professional development and other supports (e.g. curriculum coaches) to promote improved teaching. (Hamilton et al., 2008, p. 39)

Other research has reported negative effects. There are research studies that support the conclusion that large-scale assessments have dictated a pace of instruction that can preclude more open-ended exploration of topics (McMillan, 2007). An additional finding is that the content of the these assessments has resulted in a reallocation of instructional time away from non-tested areas in order to devote more instructional time to tested subjects (Hamilton et al., 2008). Another short coming of these assessments is that they identify students whose performance is sub-par relative to academic standards; however, they provide only minimal diagnostic information as to how to improve performance (Ryan, 2002).

Contemporary Issues Regarding the Summative Use of Assessment

The format of these large-scale assessments is changing and illustrates how the role of federal and the state governing bodies in setting educational policy that affect assessment practices have become intertwined. Education has generally been considered to be a state power because the language guaranteeing that all resident children receive a public education at public expense is embedded in the constitution of all 50 states (McDermott, 2011b). The Massachusetts Education Reform Act (MERA), enacted by the Massachusetts legislature in 1993, is an example of a state exercising this power in a manner that directly impacted assessment practice and led to the development of MCAS (McDermott, 2011a).

The federal government has also had a role in the development of large-scale assessment system through various legislative initiatives beginning with the Elementary and Secondary Education Act (ESEA) of 1965 (House, 1993). The 1994 reauthorization of ESEA, entitled Improving America's Schools Act (IASA) and the 2001 reauthorization, entitled No Child Left Behind (NCLB), resulted in demands for more universal testing accompanied with sanctions for chronically underperforming schools (McDermott, 2011a). Although the most recent reauthorization of ESEA has been pending for over five years, it is likely that the format of these large-scale assessments will change (Gewertz & Robelen, 2010).

Wang, Beckett, and Brown (2006) identify several ways in which the format of these assessments can be improved. The assessments can be aligned to new content standards that are more developmentally appropriate and incorporate new research in regard to cognitive plasticity at different stages of development. The panels that develop these tests can be broadened to include classroom teachers and cognitive-developmental and social psychologists. New formats can employ computerized adaptive testing that incorporate assessments that are more complex than multiple-choice questions.

These changes in format are evident in the revision that is currently underway to the large-scale assessment system in Massachusetts under the jurisdiction of The Partnership for Assessment of Readiness for College and Careers (PARCC), a consortium of states working together on this project (PARCC, 2012). The new assessment system is anchored to a new set of learning standards, the Common Core State Standards (CCSS). In 2011Massachusetts adopted their version of the CCSS in the content areas of English Language Arts and Math.

The extent of the changes in format is not fully understood at this time because the design phase began in 2010. PARCC projected a timeline of piloting the new assessments for two years beginning during the 2012-2013 school year with a projected date for the full operational administration during the 2014-2015 school year. Based on the information that is currently available from PARCC, the format of the assessment will be substantially changed. The range of assessment will encompass the entire kindergarten-grade 12 spectrum. Assessments at the kindergarten-grade 2 range will be formative in nature comprising observations, checklists, classroom activities, and protocols. Assessment at the grade 3-grade 8 range will be both summative and non-summative. There will be two types of required summative assessments: a Performance-Based Assessment (PBA) and an End-of-Year Assessment (EOY). The PBA will be administered close to the end of the year. In English Language Arts the focus will be on writing effectively when analyzing text. The focus in math will be on applying skills, concepts, and understandings to solve multi-step problems. The EOY in English Language Arts will focus on reading comprehension. The optional non-summative components will consist of a Diagnostic Assessment that will serve as an early indicator of student knowledge and skills in order to tailor instruction to meet the needs of individual students. There will also be a Mid-Year Assessment comprised of performance-based items and tasks with an emphasis on hard-to-measure standards. The assessments for high school are similar to those for grades 3-8; however, they are administered closer in time to when instruction occurs. This work is all in the very preliminary stages and not available for review by administrators working at the local level at this time.

In summary, the summative use of assessments is a long-standing practice. Research on the effects of the summative use of assessments on student learning supports findings of both positive and negative effects. It is likely that state-mandated, large-scale assessments will continue to occupy an “out-sized” place in the overall landscape of assessments (Rothman, 2010). However, the format of these assessments will likely change as a result of a new set of learning standards (PARCC, 2012).

The Evaluative Use of Assessment

When the primary purpose of assessment is the systematic utilization of data to gauge the value, effectiveness or efficiency of educational policies, programs and personnel, it can be referred to as the evaluative use of assessment (CCSSO, 2002). The rise in the use of this type of assessment is linked to the increasing public demands for accountability and outcomes-driven measures of performance. Simply stated, stakeholders want to know about the effects and merits of the programs they are being asked to fund, implement, vote for, or participate in. To understand the rationale for including assessments that are used for evaluative purposes, a brief review of the legislative initiatives that have ushered in their use is warranted.

History of Federal Legislative Initiatives

The forces that shape the contemporary landscape of assessments have roots in the Elementary and Secondary Education Act of 1965 (ESEA). Title I of ESEA aimed to improve the caliber of basic education for underserved and economically disadvantaged children by providing substantial monetary support to low-income school districts. To ensure that the Title I ESEA funds benefited the targeted population, the legislation contained a series of evaluation requirements that were limited in scope; only schools that received Title I funds had to assess their students and school officials had broad discretion to decide what they assessed and how they assessed them. With the passage of the ESEA, evaluation requirements designed to study the impact and effects of social programs (D. Campbell, 1969) and to support continuous program improvement (Patton, 2008) have become part and parcel of every federal grant since 1965 (House, 1993). The significance of this legislation is that it represents a historical legislative link between locally administered assessments and externally imposed expectations of performance.

Concern over the caliber of America’s public schools continued to increase in the ensuing years, fueled by reports including *A Nation at Risk,* published in 1983 by the National Commission on Excellence in Education. This report concluded that our national security was in peril because of substandard education in American public schools. These concerns galvanized the political and business communities in a united effort to bring about substantial reform. The nature of these efforts was characterized by a theory of action referred to as the “tight-loose coupling” principle: “establish standards (goals and standards), provide flexibility for states and local districts, then hold people accountable” (Cross, 2004, p. 91).

Over the years these efforts manifested in several initiatives, such as the 1989 Charlottesville Education Summit, the 1990 National Education Goal, and Goals 2000. Numerous groups, such as the National Council of Teachers in Mathematics, were also working to establish the first set of voluntary standards for their content areas. Several trends characterized the school reform initiatives at this time: the role of the federal government was expanding, a common agenda and set of goals was beginning to emerge, and there was a strong focus on the outcomes of the educational system.

These trends came together in the 1994 reauthorization of ESEA, referred to as the Improving America’s Schools Act, and had significant implications for the evolution of assessment practices in public schools. The elements of this legislation that affect assessment practices include the provisions that required states “to develop content and performance standards for all children and replace generic multiple choice tests for Title 1 students with ones that were aligned to the standards, creating coherence between standards and assessments” (Cross, 2004, p. 110). Through this legislation, the federal government had placed itself in the position to set the agenda for education in almost every school district and state. The era of accountability and outcome-driven education was dawning.

This legislation led some states to change their accountability policies. Although each state was at liberty to develop its own standards and unique system of assessment, most developed large-scale criterion referenced systems (Popham, 2009). Substantial costs were associated with the implementation of these assessment systems. In 1993, the General Accounting Office pegged the cost nationally of state- and district- level testing at $36 million or about a $14.50 /pupil cost (Cizek, 2007). To ensure accountability, the assessment results were widely-reported and linked to significant sanctions for students and educators. The 1994 amendments to ESEA also established that states had until the 2000 school year to implement a state accountability assessment. Although the standards-based school reform movement was already fueling change at the state level, the federally-established time line for compliance made it clear that all states had to quickly overhaul their assessment systems.

By 2001, only 17 states were reported to be in full compliance (Cross, 2004). This slow rate of progress towards comprehensive reform fueled the next phase of significant legislative action. Once again, political and social forces united and focused their efforts on the 2001 reauthorization of ESEA, commonly referred to as the No Child Left Behind Act (NCLB). The assessment-related provisions of NCLB dramatically increased the assessment obligations of school districts. NCLB required all public schools to administer assessments in reading and mathematics to all students in grades 3 through 8 and once in grades 10 through 12 no later than the 2005-2006 school year. By 2007-2008, states had to add tests in science at least once in grades 3 through 5, 6 through 9, and 10 through 12. It is estimated that these regulations affected at least 25 million students annually (Abrams, 2007). States could potentially use locally developed or off-the-shelf commercial tests to meet the testing obligations (Manna, 2004); however, the federal government exerted considerable pressure to employ traditional standardized testing instruments (Popham, 2009). The result is that most states continued to develop and utilize large-scale standardized assessment instruments.

NCLB also required all states to participate, at federal expense, in yearly testing for 4th graders in reading and 8th graders in math using the National Assessment of Educational Progress (NAEP). NAEP is a national testing program, begun in 1969, that measures broad trends in achievement for students nationwide and provides an independent source of information about how students in participating states are performing relative to the nation as a whole. The results on NAEP could now be used as a “de facto validity check on state tests” and states would now have to justify any discrepancy between any reported levels of high performance on state measures with low performance on NAEP (Manna, 2004, p. 139).

NCLB also redefined accountability. All states had to employ at least three levels of performance—advanced, proficient and basic— to their locally developed assessment. Of greater significance, NCLB introduced the concept of adequate yearly progress (AYP). AYP is a goal-setting mechanism that sets the performance target of proficiency for all students in reading and mathematics by 2014. States now had to disaggregate student test results and report them along a spectrum of measures, such as gender, socio-economic status, or status as a regular education or special education student. The intent of AYP was to ensure that schools were closing the achievement gap between advantaged students and their disadvantaged or racially and ethnically diverse peers. However, the calculation of AYP was complex and essentially a statistical impossibility. Complicating matters even further, states individually set their baseline of performance from which they demonstrated improvement (Manna, 2004).

A recent development in Massachusetts reflects the efforts of states to address some of the flaws in the use of AYP as a metric for identifying under-performing schools. Beginning in 2010 with the passage of the state’s Achievement Gap Act, Massachusetts schools operated under a dual accountability system (MA DESE, 2012). Districts and schools were assessed using the state’s five-level Framework for District and School Accountability and Assistance and also assessed using the AYP metric. In November 2011 Massachusetts applied for a waiver from the United States Department of Education (ED) claiming that NCLB’s rising targets have resulted in AYP no longer being useful in identifying schools and districts most in need of assistance and intervention. In their waiver request Massachusetts noted that by applying the AYP in 2011, 81% of all schools and 90% of all districts were designated as not making yearly progress despite the fact that Massachusetts outscored all other states on NAEP at the 4th and 8th grade levels. In February 2012, the ED granted Massachusetts flexibility to this provision. In its place, Massachusetts will maintain the state’s five-level Framework and districts will continue to be identified by their lowest performing school. The state also established the goal that by 2017 all aggregate and student subgroups will reduce by half the proficiency gap between the group’s current achievement levels and the goal of having 100% of student proficiency. They added a new student sub-group category of “high needs,” composed of students who are low-income, have a disability, and a history of limited English proficiency. Massachusetts will continue to use the Composite Performance Index (CPI) as the metric of achievement.

In summary, legislative initiatives at the federal level have fostered an increase in the evaluative use of assessment. This increase is rooted in the public's growing concern about a decline in the quality of the educational system in the United States and a desire to ensure that all students, not just advantaged students, are achieving at high levels of academic proficiency in key content areas. To meet the requirements for accountability, states mandated large-scale standardized assessments that are used for both summative and evaluative purposes. The research on the effects of these assessments on students' learning was reviewed in the previous section and supports findings of both positive and negative effects. At the federal level, legislation imposed new standards on all states but there are signs of growing flexibility.

Coherence Amongst Components

Coherence refers to the alignment of the various components of the educational process to a common point of reference. During this era of standards-based reform, the point of reference is an articulated set of academic standards (Popham, 2006). The most recent effort to refine standards has been led by the CCSCO in partnership with the National Governors Association and has culminated in the Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science and Technical Subjects (MA DESE, 2010). These new standards aim to be (1) research and evidence based, (2) aligned with college and work expectations, (3) rigorous, and (4) internationally benchmarked. The standards, which are expressed in broad terms, have to be deconstructed. To do this, educators have to "identify what the standards will look like as targets of daily instruction for the classroom teacher" (Chappuis et al., 2010, p. 55).

Once the targets of daily instruction are identified, an assessment system can be aligned to those learning targets.

Wiggins and McTighe (2007) paint a disheartening picture of the level of coherence between assessment practices and the articulated set of learning standards.

> Currently, few schools or districts have a robust assessment system that is designed from the start to align closely with standards, program goals, or long-term mission. In part, this is because few educators have been adequately trained to design valid assessments of broader, long-term goals. Moreover, the great majority of classroom- and district-level assessments tend to focus on content mastery and the lower-order of cognitive processes of Bloom's Taxonomy, not on understanding and performance on complex tasks that demand transfer. (p. 79)

They advocate for educators at the local level to begin the process of deconstructing standards by initially articulating a concept of the long-term mission of schooling. Wiggins and McTighe support the concept that the long-term mission of schooling is to "learn to use powerful ideas to make schoolwork connected and meaningful and to transfer learning thoughtfully and efficiently to novel situations and problems" (p. 13). By working backwards from that, or any other concept, educators will have a better chance of identifying meaningful targets. Once the targets are defined, then assessments can be matched to them and instructional practices can also follow suit. This is an iterative process that evolves over time.

Robust Capacity for Data Management

A robust data management system, characterized as one that can deliver information to a multitude of users in a timely manner, is a core feature of a balanced assessment system (Boudett et al., 2005; Boudett & Steele, 2007; Love et al., 2008). The need to develop the capacity for data management is reflected in the Massachusetts

Department of Secondary and Elementary Education Race to the Top plan. This plan has five objectives, one of which is to, "provide educators with the real-time, actionable data they need to meet the needs of every student" (MA DESE, 2010, p. 18). To that end, the DESE has set the goal to 1) transform the Commonwealth's data system by expanding the capability of the existing Data Warehouse and implementing the Schools Interoperability Framework to automate data uploads, 2) invest in new technology, and 3) strengthen and expand training in the use of data by developing a new series of online and in-person courses (MA DESE, 2010).

An effective data management system will have the capacity to collect, handle, and report results generated from a wide variety of educational assessments in addition to the capacity to link that data to other sources of relevant student information, such as attendance and to instructional interventions (Halverson, Grigg, Prichett, & Thomas, 2007). Advances in technology have had a positive impact on the capacity of systems to handle large and complex data sets and have also enabled new types of assessments, such as computer-adaptive testing that adjusts to the pattern of response of an individual student and can yield a more accurate measure of achievement levels (Rothman, 2010). The development and maintenance of these complex systems necessitates a level of technical expertise at both the state and local district level (Lasky, Schaffer, & Hopkins, 2009).

The contribution of an effective data management system to a balanced system of assessment is also dependent on a different type of capacity—the capacity to make meaning of the data. As part of developing this capacity, Love et al. (2008) highlight the need to foster a level of awareness on the part of educators who use data.

> Data have no meaning. Meaning is imposed through interpretation. Frames of reference, the way we see the world, influence the meaning we derive from data. Effective data users become aware of and critically examine their frames of reference and assumptions. Conversely, data users themselves can also be a catalyst to questioning assumptions and changing practices based on new ways of thinking. (p. 5)

The capacity to make meaning from data can be enhanced by practices, such as data teams that engage in collaborative inquiry (Boudett et al., 2005; Boudett & Steele, 2007; Earl & Timperley, 2009; Love et al., 2008). Highly functioning data teams engage in activities, such as building assessment literacy, creating data overviews, and examining the link between student improvement and instructional practice. Data retreats can afford a mechanism for educators to make sense of achievement data and chart plans for instructional change (Sargent, 2003). Professional learning communities is another mechanism for building capacity where teams of educators work together to develop competencies in this area (DuFour, Eaker, & Dufour, 2005).

In summary, a robust capacity for data management implies more than just the capacity to handle the technical manipulation of data. An effective data management system must also include features that foster the capacity of educators to meaningfully and thoughtfully interpret the data. Practices, such as data teams, data retreats, and on-going work through professional learning communities, can foster competencies in this area.

High-Quality and Diverse Assessments

A balanced system of assessments will incorporate high-quality and diverse measures. High-quality assessments are characterized by strong psychometric properties. In the case of the large-scale mandated standardized assessments that are part of a state's accountability system, the caliber of these measures is especially important because of the high stakes consequences that are linked to the results (Pellegrino et al., 2001). By 2002, when states began to develop these systems, there was "almost no literature on the validity of accountability systems" (CCSSO, 2002, p. 38). To address this shortcoming numerous agencies, including The National Center for Research of Evaluation, Standards, and Student Testing, The National Center for Improvement of Educational Assessment, The Division of State Services and Technical Assistance of the Council of Chief State School Officers, and The State Collaborative on Assessment and Student Standards, have provided guidance and technical support to state officials to design and refine their state accountability systems (CCSSO, 2004).

In the case of smaller-scale assessments which are typically administered more frequently, educators cannot assume that the assessments have been designed with a similar level of vigilance to psychometric properties. In many schools, teachers may need to, want to or are expected to create their own assessments; however, they typically lack the expertise to evaluate the adequacy of these tests on their own (Love et al., 2008). This situation highlights the need for the research community to undertake the task of establishing the technical adequacy of measures, other than large-scale standardized assessments, that can be administered by teachers in their classroom.

The development of curriculum-based measures (CBM) provides a good example of how this was accomplished with an assessment that is typically used for formative purposes. CBM refers to measurement activities that use direct observation and recording of a student's performance with material from the local curriculum as a basis for informing instructional decisions (Hintze, Christ, & Methe, 2006). Since its conception in the 1980s, researchers in this area have worked to both establish the technical adequacy of these measures for reading and math (Clarke & Shinn, 2004; Deno, 1985; Fuchs & Fuchs, 1993; Thurber, Shinn, & Smolkowski, 2002) and to study the effects on student achievement (Capizzi & Fuchs, 2005; Fuchs, Fuchs, Hamlett, & Stecker, 1991).

Diverse assessments imply that there is a variety of measures such that students from a wide range of ethnic and socio-economic backgrounds can participate equally (Haui et al., 2006). The need for a diverse range of assessments is greater today than in the past due to the changing composition of our nation's student population (Durden, 2008). From 1979 to 2005 the number of children, ages 5-17 years, who enrolled in public school and spoke a language other than English increased from 3.8 million to 10.6 million. Ethnic diversity also increased with the highest concentration of these students clustered in high-poverty schools where more than 75% of students qualified for free or reduced-fee lunch. To address the needs of culturally diverse students in regard to assessment, the Association for Assessment in Counseling (AAC) (2003) articulated over 68 standards in their policy statement, *Standards for Multicultural Assessment*. The ACC advocates that

> test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain. (p. 3)

The need to incorporate culturally-sensitive assessments will increase as the diversity of our population increases.

Ensuring a Minimum Burden

The burden of incorporating assessment into educational practice needs to be taken into account. All told, the development, implementation and on-going application of a balanced assessment system necessitates a significant allocation of a school district's resources in terms of time and money. Given that there is always competition for limited resources, a balanced system of assessments should be designed such that it places a minimum burden on students and staff to obtain, analyze, interpret and use assessment information (Boudett & Steele, 2007). The strategies to streamline an assessment system are similar to other efforts to reduce redundancies in other areas of an operating system.

Articulating a clear vision of the role of assessment is an important step in creating an efficient system of assessments (Wiggins & McTighe, 2007). This entails articulating how assessments will be used for formative, summative and evaluative purposes and will need to incorporate the realities of state-mandated practices, such as large-scale standardized measures. The vision should incorporate the needs of the different constituents, ranging from administrators, staff, parents and students, and their unique need for different types of assessment data.

Conducting an audit of current assessment practice is another strategy (Chappuis et al., 2010). An audit typically consists of gathering information about critical features of each assessment, such as timing of administration, connection to the standards and learning targets, targeted grade level, data management requirements, intended purpose, primary users of results and key decisions, that the results will inform. Maintaining this audit in electronic form is a strategy that enables school administrators to view this work as a "living document" that can be updated frequently(Boudett & Steele, 2007, p. 16).

Teacher Competency in the Educational Assessment of Students

There is a growing recognition that administrators and classroom teachers have to be assessment literate, implying that they have the necessary skills to ethically and appropriately develop, administer, score, interpret and use assessment measures (Popham, 2009). There is also a growing recognition that many practicing teachers have not acquired this set of skills (Mertler, 2003; Popham, 2006). This gap in skills has been attributed to an historical lack of adequate preparation in teacher pre-service programs (Mertler & Campbell, 2005; Popham, 2009; Vogel, Rau, Baker, & Ashby, 2006). The majority of teachers in today's classrooms completed their teacher training program when "there was no requirement that they learn anything about educational assessment" (Popham, 2009, p. 5).

In 1987 the American Federation of Teachers (AFT), the National Council on Measurement in Education (NCME) and the National Education Association (NEA) established a joint task force whose work culminated in 1990 in *The Standards for Teacher Competence in the Educational Assessment of Students*, in which standards were articulated (AFT et al., 1990). Their concept of the role of assessment was broad and incorporated the formative, summative, and evaluative use of assessment with the specific goals of giving feedback to a student about his or her progress, judging instructional effectiveness, and informing policy. They state:

> (1) teachers should be skilled in choosing assessment methods appropriate for instructional decisions
> (2) teachers should be skilled in developing assessment methods appropriate for instructional decisions
> (3) teachers should be skilled in administering, scoring and interpreting the results of both externally produced and teacher produced assessment methods

> (4) teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement
> (5) teachers should be skilled in developing valid pupil grading procedures which use pupil assessments
> (6) teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators
> (7) teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information. (AFT et al., 1990)

The task force intended their standards to be used by teacher educators in the design of pre-service training programs, by teachers in their self-assessment and professional development plans, by workshop instructors in their design of professional development trainings for pre-service teachers, and as an impetus to measurement specialists and teacher trainers to adopt a broader conceptualization of student assessment. Although dated, the standards articulated by the AFT, NCME, and NEA continue to be the reference point for many research studies on assessment literacy for classroom teachers (Chen, 2005; McMillan, 2007; Mertler & Campbell, 2005; Plake & Impara, 1993).

Research efforts have focused on how to translate these standards into measurable terms to gauge the level of assessment literacy of pre-service and in-service teachers (Mertler & Campbell, 2005; Plake & Impara, 1993). Plake and Impara used *The Standards for Teacher Competence in the Educational Assessment of Students* (AFT et al., 1990) to develop their *The Teacher Assessment Literacy Questionnaire*. This instrument consists of 35 unrelated items in a multiple choice format. Plake and Impara surveyed a representative sample of educators from 98 different school districts in 45 states that yielded a return of 555 respondents. On average, respondents answered 23 out of 35 questions correctly. Respondents performed the highest on the standard related to the administration, scoring and interpreting assessment results and the lowest on the standard related to communicating assessment results. Teachers with some literacy training scored higher than those with no training. The reliability for the entire test (KR20) was .54. C. Campbell, Murphy, and Holt (2002) and Mertler (2003) used *The Teacher Assessment Literacy Questionnaire* with pre-service and in-service teachers with comparable findings, however, they each separately recommended a revision of the instrument based on their experience that the questionnaire was "difficult to read, extremely lengthy, and contained items that were presented in a decontextualized way" (Mertler & Campbell, 2005, p. 9).

In 2003, Mertler and Campbell developed the *Assessment Literacy Inventory* (ALI), which is also based on used *The Standards for Teacher Competence in the Educational Assessment of Students* (AFT et al., 1990) and tested its psychometric properties. This inventory consists of five classroom related scenarios, each of which is followed by seven questions that correspond to the seven standards. They conducted a two-stage pilot with 152 pre-service teachers in Fall 2003 and 249 pre-service teachers in Spring 2004. After revisions to specific items, their item analysis of their final version yielded an overall instrument reliability (KR20) of .74. Based on this analysis, they concluded that the ALI provided a practical mechanism for measuring assessment literacy (Mertler & Campbell, 2005). Since their work had been exclusively with pre-service teachers, they recommended further studies be undertaken with in-service teachers to determine the appropriateness of *ALI* as a measure of assessment literacy with this population.

Additional research has also focused of the efficacy of in-service training programs to increase the assessment literacy skills of practicing teachers (Lukin, Bandalos, Eckhout, & Michelson, 2004; Vogel et al., 2006). Lukin et al. reported on the efforts in Nebraska to provide in-service training aimed at developing the assessment literacy of practicing teachers. In lieu of designing a statewide assessment system, Nebraska opted for

> the development and implementation of a statewide system of district-level assessments as a means of holding districts accountable for maintaining a rigorous curriculum while at the same time maximizing student achievement through improvements in classroom assessment practices. (p. 26)

To foster the assessment literacy of staff and to enable them to develop district-level assessments, three separate training options were offered by different entities within Nebraska. In the first option, the University of Nebraska-Lincoln developed the National Assessment Cohort (NAC), a formal course of study consisting of 18 hours of graduate level work that qualified participants as an assessment resource teacher. The goal of this course of study was to develop the knowledge and skills necessary for the development and implementation of both classroom-level and district-level assessments. Participants who completed this course of study generally reported an increased confidence in their knowledge and skills in a variety of assessment related areas. This initial feedback led to the recommendation to include more teachers in these training models and even expand it to include building-level administrators.

The second training option entailed the state of Nebraska contracting with the Assessment Training Institute (ATI) with four schools participating in the initial study. Local administrators used materials developed by Stiggins at ATI and organized Assessment Literacy Learning Teams (ALLT) consisting of teachers and administrators. The goal of this training was to develop literacy skills related to classroom assessments. Participants reported an increase in confidence and skill in the area of assessment and there was also evidence of positive outcomes for students, most notably an increase in achievement levels on the Metropolitan Achievement Test, a district-level measure.

The third option was an outgrowth of ALLT and supported through a partnership between the University of Nebraska-Lincoln and the Lincoln Public Schools and was referred to as the Pre-service Assessment Literacy Study Group and In-service and Pre-service Assessment Literacy Study Group. In this model, pre-service and practicing teachers studied together in a learning team format over the course of one year. This model generally received positive reviews, but insufficient data prohibited the adequate analysis of its effectiveness.

Overall, the researchers concluded that these three models of in-service training offered promise; however, they noted that developing assessment literacy requires a long-term commitment. They identified several features of these training models that appeared to be related to positive outcomes. Key features included the flexibility to adapt the training to local conditions and the use of a learning team format.

Vogel et al. (2006) reported on four reform initiatives in Illinois that also incorporated professional development programs as the means to increase assessment literacy. Each initiative had the goal of developing assessment literacy skills through in-service training, but each initiative developed a different model of professional development training. The last and most successful of the reform efforts, referred to as The Standards-Aligned Classroom Initiative, was developed by an intermediate government body, comprising primarily publicly elected regional school superintendents. Working with an outside consultant, this group designed a three-year professional development program that was piloted in over 200 schools. The first year of the program concentrated on developing assessment literacy and instructional philosophy. Participating personnel, consisting of teachers and administrators, met as learning teams every two weeks for at least 1.5 hours to read and discuss *Student-Involved Classroom Assessment* (Stiggins, 2001), and they presented at a "share fair" with other teams from around the state. The second year focused on application in the classroom. Returning participants attended a 1-day training session to refresh the concepts developed in the first year and then continued to meet in learning teams at least four times throughout the year to create lesson plans and study additional material. Presentation by the team at a statewide "share fair" was also expected. The third year focused on sustaining the work and broadening it beyond the learning teams. Returning teams were asked to mentor teams that were just starting the process. And, as in previous years, teams created presentations for a statewide "share fair."

This model of professional development was evaluated by external consultants. The average total effect size for the 404 first-year participants was 1.02 with reported increase in familiarity with standards and application of the standards to instruction. The average total effect size for 287 second-year participants was .59. By comparison, "effect sizes of .20 or greater have been considered substantial for training programs in education " (Vogel et al., 2006, p. 51). Based on this positive effect and some indications of positive effect on student achievement scores, the recommendation to the Illinois State Board of Education was to mandate this model of training in low-performing schools throughout the state.

In summary, teachers have historically been ill-prepared in the educational assessment of students. The AFT, NCME and the AFT jointly articulated *The Standards* *for Teacher Competence in the Educational Assessment of Students* in 1990 and, although dated, these standards continue to serve as a benchmark for performance. Research on assessment literacy has focused on developing instruments to gauge the current level of competence of pre-service teachers. Research on the effect of in-service professional training programs on assessment literacy has encompassed a variety of training models and has supported the finding that these trainings have a positive effect on the levels of competency in the educational assessment of students.

This review of the literature underscores the challenges faced by educators and policy makers as they strive to design educational assessments that genuinely enhance student learning. One area of study has focused on the development of a balanced system of educational assessment that is considered to have a comprehensive range of assessments used for formative, summative and evaluative purposes. The assessments all cohere to a single set of standards, are of high-quality and diverse enough to fairly assess a wide-range of students. A balanced system has a robust capacity for data management and imposes a minimum burden on staff to develop, obtain, analyze, interpret and use assessment information. Despite the potential of a balanced system to harness the power of assessment, it is not common practice. The successful deployment of these systems is highly dependent on the level of competency of staff in the educational assessment of students.

CHAPTER 3

METHOD

The purpose of this study is to gather, analyze, and report data about the current status of one school district's pre-kindergarten through 8th grade system of math assessment and the level of assessment literacy of staff at these grades levels. To conduct this study, I adopted the framework of a utilization-focused evaluation (UFE). Working collaboratively with the administrators, we identified a core set of three research questions that guide this study. The nature of the research questions dictates a mixed methods approach.

Design

Evaluation is a demonstrated method for analysis and a means of building capacity (Smith & Freeman, 2002). There are numerous types of evaluations, including those that are characterized as responsive, goal-free, consumer-oriented, theory-driven, and utilization-focused (Russ-Eft & Preskill, 2009). The characteristics of a UFE make it a suitable choice for this study.

A UFE is distinct from other types of evaluation by the extent to which the evaluation is tailored to the unique needs and interests of the participants and by the emphasis that is placed on the specific use of the findings. Patton (2008), considered the founding father of UFE, explains these unique features of a UFE.

> Program evaluation is the systematic collection of information about the activities, characteristics, and results of programs to make judgments about the program, improve or further develop program effectiveness, inform decisions about future programming, and/or increase understanding. *Utilization-focused program evaluation* is done for and with specific intended primary users for specific, intended uses. (p. 39)

Over the course of several months and several planning sessions, I developed the plan for this UFE through extensive collaboration with the administrators in this school district. The administrators narrowed the focus to their math assessment system in grades pre-kindergarten through 8th grade range. The rationale for their decision is based in the unique demographic characteristics of their student population and their curricular initiatives at the time of this study.

Another feature of a UFE is the extent to which there is an emphasis on the use of the findings. Utility is at the core of a utilization-focused evaluation. As Patton (2008), explains,

> *Utilization-focused evaluation* begins with the premise that evaluations should be judged by their utility and actual use; therefore, evaluators should facilitate the evaluation process and design any evaluation with careful consideration for how everything that is done, from beginning to end, will affect use. Use concerns how real people in the real world apply evaluation findings and experience the evaluation process. (p. 37)

To incorporate aspects of use early on in the planning these administrators identified that they could use the findings to redesign their math assessment practices, to reallocate district resources, and to plan for in-service professional development in the area of assessment literacy.

The nature of the research questions also dictates a mixed methods approach. A mixed methods approach refers to collecting and analyzing both qualitative and quantitative data in a single study (Creswell, 2003). In this study the administrative team wanted data about the level of assessment literacy of staff. This information was obtained through the application of quantitative methods. They also wanted to understand the perspective of specific classroom teachers as they employ assessment practices in their classroom. This type of information is best gleaned through the use of qualitative methods. By merging the two methods, there is the potential for a better-informed set of findings than would be possible through the use of either a quantitative or qualitative in isolation.

Evaluation Research Questions

I collaborated with the administrative team to develop the research questions. The questions reflect the unique needs of the district at the time of the study. The scope of the questions takes into account what is realistically feasible for a sole researcher. The research questions are as follows:

Question #1: To what extent do we currently have a balanced and comprehensive system of math assessments in pre-kindergarten through 8th grade?

Question #2: To what extent are our 6th and 7th grade teachers using math assessment to facilitate the transition of continuing and in-coming students into 7th grade?

Question #3: What is the level of competency of our staff relative to established standards of competency for the educational assessment of students?

The administrative team identified math as the focus of this research study based on their analysis of student performance levels on the Massachusetts Comprehensive Assessment System (MCAS) from spring 2010. They were concerned that their students in the aggregate performed at lower levels in math than in English Language Arts. They had also been less successful in improving student performance levels in math over the years of MCAS testing. See Table 2 for a summary by school of their 2010 MCAS

Composite Performance Index (CPI), a measure of aggregate student performance with a potential score of 100, and the accompanying descriptive rating.

Table 2
Composite Performance Index and Performance Rating on 2010 Math MCAS

School	CPI	Rating
Elementary School A	75	Moderate
Elementary School B	83.1	High
Elementary School C	77.5	Moderate
Elementary School D	76.6	Moderate
Middle/High School	76.3	Moderate

The specific methods for gathering and analyzing the data are presented on a question-by-question basis. A timeline provides an overview of the entire process and highlights the intervals at which data will be gathered and findings shared with the key stakeholders.

Stakeholders

The primary stakeholders for this study are the district administrators, including the Superintendent, Assistant Superintendent, Elementary Curriculum Coordinator, Technology Coordinator, Director of Special Education and 4 out of the 5 building Principals. One principal obtained employment in a different district during the course of this study and the replacement was too unfamiliar with district practices to meaningfully participate. Demographic information, consisting of their role in the district and the number of years in their current capacity, was gathered for all administrators. Given the small number of administrators, their confidentiality is preserved by only identifying roles as to whether it is a Central Office (CO) or Building Level (BL) position because that distinction is relevant to the interpretation of the results. Building-level administrators will typically have a more thorough understanding of the implementation of practices whereas as a Central Office-level administrator will typically have a more thorough understanding of the scope of practice throughout the district. The demographic information on the administrative team is summarized in Table 3. Consent for voluntary participation was obtained from all administrators. See Appendix A for consent letter.

Table 3
Demographic Information on Administrative Team

Administrator	**Number of years employed in the district in this capacity**
CO1	4 years
CO2	3 years
CO3	3 years
CO4	4 years
CO5	3 years
BL1	5 years
BL2	4 years
BL3	8 years
BL4	6 years

Setting

The regional school district that is the setting of this study is located in a rural area in Massachusetts. Although isolated from major metropolitan areas, the district is culturally diverse due to its proximity to several colleges, universities and private high schools. The school district is under the jurisdiction of one school committee whose members are elected from four participating towns.

The student population is housed in four elementary schools, each serving a different municipality, and one regional middle and high school that serves the four-town region. All elementary schools have a similar grade configuration of pre-kindergarten through 6th grade. The elementary schools differ significantly from one another in terms of enrollment and range from a low of 58 students to a high of 294. Despite the small numbers of students in some schools, each community is committed to retaining its own elementary school. The regional middle/high school is located on one campus and serves all students from 7th through 12th grade.

All schools offer the option for parents from out-of district to enroll their children at all grade levels. Currently the district has substantially more students who choice-in than choice-out of the district. A major point of entry for students to choice-in is at 7th grade. On average, approximately 20% of the 7th grade class comprises students who are entering the district for the first time and coming from either out-of-state or other Massachusetts communities.

The number of faculty at each school varies in relation to the size of the student body. We decided to invite only the general education and special education teachers in grades pre-kindergarten through 8th grade to participate. The rationale for that decision relates to the topic of the study; these are the staff members who are primarily responsible for administering math assessments. The total number of students along with the combined total of general education and special education teachers at each school for the 2010-11 is summarized in Table 4. Consent for voluntary participation was obtained from all participating teachers. See Appendix A for consent letter.

Table 4
Total Number of Students and Faculty at Each School in 2010-2011

School	Total number of students	Combined total of faculty
School A	53	4
School B	54	5
School C	213	11
School D	276	13
Middle	208	9

Procedures for Question #1

The intent of Question #1 was to gather comprehensive data in regard to the various math assessments that the district currently uses. I used a combination of qualitative methods to gather data to answer this question. Specific methods included a semi-structured interview that I designed specifically for this study, a survey of school and district-level assessment measures developed by the Pearson Assessment Training Institute Staff and a survey of classroom-level assessment measures that I designed for this study.

Semi-structured Interview

I conducted an individual semi-structured interview with each Central Office and Building-level administrator. I invited all nine of them to participate and all nine accepted. The use of a semi-structured interview was particularly well-suited to this study in contrast to alternative approaches, such as unstructured, structured or focus group interviews. A semi-structured interview consists of a set of questions that guide the interviewer to ensure a level of consistency while also providing some flexibility to gather information tailored to specific individuals through additional probing (Russ-Eft &

Preskill, 2009). This approach enabled me to gather some consistent information while also allowing for unique participant perspectives and experiences to emerge.

Prior to conducting the interviews, I piloted the core set of interview questions on a building-level administrator in a different school district to ensure clarity of phrasing. The core set of questions is listed in Appendix B. I audio-recorded each interview and then personally transcribed them. The interviews lasted about 50 minutes, on average. All recordings and transcribed material will be destroyed upon completion and acceptance of this dissertation.

I analyzed the transcripts of the interviews and coded them for thematic trends. Coding involves a thorough analysis of the text from the transcribed interviews to identify key themes that can be labeled and categorized (Russ-Eft & Preskill, 2009). The development of the themes entailed both an open-coding in which the themes emerged from the data and a constant comparison method in which the themes emerged from the on-going comparison of the data with the theoretical literature on assessment.

Survey of School and District Level Assessment Measures

The Assessment System Self-Evaluation (ASSE), developed by the Pearson Assessment Training Institute, is a survey that provides a qualitative measure of the status of implementation of various assessment practices that are relevant to the development of a balanced assessment system (Chappuis et al., 2010). The ASSE identifies seven main action steps and under each action step there are sub-steps. The main action steps are summarized in Table 5. See Appendix C for the ASSE protocol.

Table 5
Assessment System Self-Evaluation Action Steps

Action step	**Description**
1	Balance the district's assessment system to meet all key users' needs
2	Refine achievement standards to reflect clear and appropriate expectations at all levels
3	Ensure assessment quality in all contexts to support good decision-making
4	Help learners become assessors by using assessment for classroom learning in the classroom
5	Build communication systems to support and report student learning
6	Motivate students with learning success
7	Provide professional development needed to ensure assessment literacy throughout the system

I chose the ASSE for several reasons. The ASSE uses constructs of a balanced assessment system that overlap with the constructs of my model of a balanced assessment system. The ASSE has recently been revised and incorporates contemporary developments in the types and use of assessment. The ASSE logically translates into action. It is just one component of a broader action guide and the administrators can easily tie it into other components. The ASSE is already in the public domain and that makes it easier to replicate its use by other administrators or researchers. A disadvantage of the ASSE is that its psychometric properties have not been established.

To tailor the ASSE more closely to the purposes of this utilization-focused evaluation, I asked the administrative team to consider only math assessment practices as their point of reference. I disseminated the ASSE to all administrators at the close of the semi-structured interview along with a stamped envelope to enable them to mail it back to me anonymously. To complete the ASSE the administrators rated each statement on a scale ranging from 1-5 with 1 corresponding to "getting started," 3 corresponding to "progressing" and 5 corresponding to "implemented." The ratings of 2 and 4 were untitled but clearly represented midpoints. I asked the administrators to select one point to represent their perspective on the status of implementation across the entire district. This point of reference was feasible because the district is small enough to enable frequent contact between building-level administrators that fosters an awareness of assessment practices in different buildings. Out of a potential of 9 respondents, 7 returned the survey resulting in a 77% rate of participation.

Survey of Classroom-Level Math Assessment Practices

I developed the survey of Classroom-Level Math Assessment Practices (CLMAP) to gather data about the types of assessments that teachers are using in their classrooms. The survey is divided into sections using the categories of formative, summative and evaluative assessments. To complete the CLMAP, respondents identified an assessment that they use and then rated how useful it is to them using a scale of 1-3 with 1 corresponding to “not helpful,” 2 to “somewhat helpful,” and 3 to “very helpful.” I provided a definition and example for each of the categories. To tailor the CLMAP to the research question, I asked respondents to report on only the math assessments. See Appendix D for the Survey of Math Assessment Practices.

Although I piloted the CLMAP with a group of randomly selected group of teachers in a different district, my review of the returned surveys revealed flaws in the design that limited my analysis. The primary flaw was that respondents identified an assessment but did not rate its usefulness despite being asked to do so in the directions. To report on the use of all assessment in my analysis, I had to create a new status of unrated that did not appear in the original survey. The other complicating factor is that the staff categorized the assessments in categories where they do not logically fit. For example, some teachers identified the Massachusetts Comprehensive Assessment System (MCAS) as a formative assessment which rarely, if ever, is the case. I decided to report all assessment practices as the teachers did regardless of the confusions as this sheds light on the current level of assessment literacy of the staff.

I distributed the CLMAP to all general and special education teachers in pre-kindergarten through 8th grade during faculty meetings in all but one of the schools. Each building-level principal had generously dedicated the majority of the meeting time to me for the purposes of this study. I distributed this survey at the same time as *The Assessment Literacy Inventory* (ALI) and most teachers chose to complete the ALI before completing the CLMAP. Consequently, some teachers did not have time to complete both measures during the timeframe of the meeting. Although some teachers chose to stay beyond the contractual timeframe of their faculty meeting to complete all surveys, others left with the survey in hand. I requested that they anonymously return the completed surveys to the school secretary. Because of scheduling conflicts at one school, I had to leave the ALI and the CLMAP with the building Principal for teachers to complete at their convenience and return to me in accompanying stamped envelopes. By the end of the study, 19 teachers out of a potential pool of 42 teachers returned the CLMAP resulting in a 45% rate of participation.

I analyzed the data from the survey by the categories of formative, summative and evaluative. Within each category I analyzed the results along two dimensions. The first dimension reports on the number of teachers who reported using the assessment. This highlights the assessments that are in common use. The second dimension reports the teachers' ratings of usefulness.

Procedures for Question #2

The intent of Question #2 was to gather in-depth information on the math assessment practices of 6th and 7th grade teachers and to understand how they use the assessments to facilitate the transition of students into 7th grade. The rationale for this question is rooted in the unique demographic circumstances of this district. Although this question is of particular importance to this administrative team, my ability to gather sufficient data had some inherent limitations. One limiting factor was the small pool of potential participants, numbering five if all chose to participate. By the end of the study only two teachers participated in the study, significantly affecting the findings.

The primary source of data for this question was the semi-structured interviews that I conducted with participants; however, I also gleaned valuable information from the interviews with administrators. I developed two different sets of questions for teachers at each of the grade levels and piloted them with a 6th grade classroom teacher in a different district to ensure clarity. The questions for the interviews are listed in Appendix E.

After receiving permission from the superintendent to proceed, I contacted each potential participant via the district e-mail to explain their role in the study and to request that they participate in a face-to-face interview. When I contacted all five potential participants in June 2011, only one participant responded positively. Consent for voluntary participation was obtained using the consent form in Appendix A. I conducted a face-to-face interview with that individual, audio-taped the conversation, and then personally transcribed them. All materials will be destroyed upon successful completion of this dissertation.

In an effort to increase the number of participants, I proposed a new approach to the Superintendent and received permission to proceed. Over the summer, I contacted the remaining teachers by mail and offered an additional incentive of a $25 gift certificate to a popular book store. This failed to recruit any new participants. In October of the 2011 I again approached the Superintendent with a third proposal. I proposed to e-mail the interview questions to the remaining pool of potential participants and to offer a face-to-face interview or the option of a written response to the questions. After receiving permission to proceed, I e-mailed all remaining participants. As a result of this third solicitation, I received a written response via e-mail from one additional teacher. I made no further attempts to solicit participation. By the end of the study, only 2 out of the 5 teachers participated.

I appreciate the time these individuals took to participate. I analyzed and reported their responses for relevant themes and insights that relate to Question #2. However, the very small number of participants generated insufficient data to support conclusions in regard to this question.

Procedures for Question #3

The intent of Question #3 was to gather data about the levels of competency in the educational assessment of students for the districts' teachers in pre-kindergarten through 8th grade. I used *The Standards for Teacher Competence in the Educational Assessment* *of Students* (AFT et al., 1990) as my point of reference for defining competency. These standards are summarized in Table 5.

To assess the current level of competency in relation to these standards, I chose to use an existing instrument, the *Assessment Literacy Inventory* (ALI). Mertler and Campbell (2005) developed this instrument for their use in prior studies and established its psychometric properties. When used as a measure of assessment literacy, the inventory had an overall reliability (KR20) of .74. The ALI is not in the public domain. I contacted both authors after an internet search that led me to their current academic affiliations. Dr. Craig Mertler, on behalf of both authors, granted me permission to use the ALI in my study.

Table 6
Standards for Teacher Competency in the Educational Assessment of Students

Standard	**Description of competency**
1	Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.
2	Teachers should be skilled in developing assessment methods appropriate for instructional decisions.
3	The teacher should be skilled in administering, scoring and interpreting the results of both externally-produced and teacher-produced assessments methods.
4	The teacher should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.
5	Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.
6	Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.
7	Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

Assessment Literacy Inventory

The ALI consists of five scenarios that depict assessments that classroom teachers typically employ in their classrooms. Each scenario is followed by seven questions, each of which corresponds to one of the seven standards of competency. Respondents have to select one answer from four possible choices that are presented in a multiple-choice format. See Appendix F for this survey.

I distributed the ALI to all general and special education teachers in pre-kindergarten through 8th grade during faculty meetings at all but one of the schools. Each building-level principal had generously dedicated the majority of the meeting time to me for the purposes of this study. I distributed the ALI at the same as the CLMAP. For purposes of this research I decided to administer this inventory on an anonymous basis in order to maximize the rate of participation, however, for purposes of planning professional development the administrative team may find it more valuable to gather data in the future so that they can identify individuals by name.

I invited all general and special education teachers from pre-kindergarten to 8th grade to participate. Members of the administrative team also expressed an interest in completing the ALI and including their results in the analysis. Consequently, there was a pool of 51 potential respondents. A total of 40 participants returned the survey, resulting in a 74% rate of participation.

Limitations of This Study

There are several limitations to this study that are important to highlight. The scope of the question was limited to what a sole researcher could investigate in a reasonable timeframe. The small pool of participants, especially of 6th and 7th grade teachers, negatively affected the ability to have valid findings. The format of the CLMAP was confusing to staff and resulted in incomplete data.

Implementation Timeline

I conducted this research over a period of months. In the initial stage of planning this study, I met with the administrative team on several occasions, beginning in October 2010. I distributed the ALI and the CLMAP in May and June of 2011 to all general and special education teachers. I conducted the semi-structured interviews with administration and one faculty member over the course of several months from June through November of 2011 and distributed the ASSE at the time of the interviews.

A key feature of a UFE is the emphasis on actionable findings. I plan on sharing my findings with the administrative team in written format and by presenting at their administrative planning meetings. The implementation timetable is outlined in Table 6.

Table 7
Implementation Timetables

Research Question #1

	Data Collection		Data Analysis	
	When	**How**	**When**	**How**
Semi-structured interviews with administrators	6/11-11/11	Individual meetings	11/11-2/12	Thematic Coding
ASSE	6/11-11/11	Individual meetings	11/11-2/12	Statistical Analysis
CLMAP	5/11-6/11	Faculty meetings	11/11-2/12	Statistical Analysis

Research Question #2

	Data Collection		Data Analysis	
	When	**How**	**When**	**How**
Semi-structured interview with teachers	6/11-11/11	Individual Meetings and via e-mail	11/11-2/12	Summary analysis

Research Question #3

	Data Collection		Data Analysis	
	When	**How**	**When**	**How**
ALI	5/11-6/11	Faculty meetings	11/11-2/12	Statistical analysis i

CHAPTER 4

RESULTS

Introduction

This utilization-focused evaluation incorporated both qualitative and quantitative methods to address the research questions. The quantitative results were generated by the *Assessment Literacy Inventory* and provide a measure of the current levels of teacher and administrator competency in the educational assessment of students. The qualitative results were generated by semi-structured individual interviews, the Assessment System Self-Evaluation, a survey developed by the Pearson Assessment Training Institute, and the Classroom-Level Math Assessment Practices, a survey that I developed for this research study. In this chapter I analyze the data from each measure individually. In the subsequent chapter I integrate them in relation to one another and to the research questions.

Survey of Classroom-Level Math Assessment Practices

The CLMAP generates information about the types of math assessments that classroom and special education teachers are currently using. Respondents identified assessments and categorized them according to their use as formative, summative or evaluative. They also rated each assessment as to how useful it is to them on a scale of 1 to 3 with 1 corresponding to “not helpful,” 2 corresponding to “somewhat helpful,” and 3 corresponding to “very helpful.”

For each category of assessment, I analyzed the teachers’ responses along two dimensions. The first dimension is a basic count of the number of teachers who report using the assessment to highlight the assessments that are in common use. It is important to note that MCAS is only administered to students beginning in 3rd grade. The *Measures of Academic Progress* (MAP), a commercially available interim-benchmark assessment developed by the Northwest Evaluation Association, has only been used for two years at the 7th and 8th grade level and for only one year at the 5th and 6th grade level. Since fewer teachers have these assessments in their repertoire, they may be under-represented in the overall count.

The second dimension focuses on the usefulness of the assessment practice. A complicating factor in analyzing this dimension is that some respondents identified a practice but did not rate its usefulness. To some extent, this lack of an assigned rating was a consequence of a flawed design of the CLMAP. In order to not lose potentially relevant information, I had to incorporate a new category of “unrated” as one of the possible rankings.

The CLMAP defined formative assessments as those assessments that “provide continuous feedback during the teaching-learning cycle with the goal of modifying instruction.” I provided the examples of non-graded quizzes and fluency measures to illustrate this type of measure. The results for formative assessment practices are reported in Table 7.

The CLMAP defined a summative assessment as those assessments that “document learning at the end of the teaching-learning cycle with the goal of documenting a level of mastery.” Examples of MCAS, quizzes and tests were provided to illustrate this type of assessment. The results for summative assessment practices are reported in Table 9.

Table 8
Formative Assessment Practices Reported in CLMAP

Assessment	# of teachers who reported use	# of teachers who assigned this rating			
		Very	Somewhat	Not useful	Unrated
Tests from commercial curriculum material	6	2			4
Quizzes	5	3		1	1
Teacher observations	5	3	1		1
Review of daily work	5	4	1		
Conferencing with student	5	2	2		1
Review of work samples	4	1			3
Homework	4	2			2
Anecdotal notes	2	2			
Pre-tests	1				1
Oral presentations	1				1
Informal self-assessment	1				1
MCAS	1			1	
MAP	1	1			

Table 9
Summative Assessment Practices Reported in CLMAP

Assessment	# of teachers who reported use	Rating of usefulness			
		Very	Somewhat	Not useful	Unrated
Tests from commercial curriculum material	14	6	4		4
MCAS	13	3	8	3	1
Teacher-made tests/exams	8	1	3		4
Quizzes	6	3			3
MAP	3	2	1		
MCAS-released questions	1				1
End-of-unit projects	1				1
Performance portfolios	1				1
Performance activity	1				1
Independent project	1		1		
Fluency measures (Mad Minute & Fast Math)	1		1		

Evaluative assessments were defined as the “systemic use of assessment to gauge the value, effectiveness or efficiency of an educational program.” The example of using

MCAS results to measure the effectiveness of the curriculum illustrated this type of assessment. Although classroom teachers are typically not involved in this use of assessment at the classroom level, their input was still solicited. The results for evaluative assessments are reported in Table 10.

Table 10
Evaluative Assessment Practices Reported in CLMAP

Assessment	# of teachers who reported use	Rating of usefulness			
		Very	Somewhat	Not useful	Unrated
MCAS	8		5	1	2
Teacher-made tests/exams	3	1	1		3
MAP	2	1	1		
Quizzes	1				1
Tests from commercial curriculum materials	1	1			
MCAS-related questions	1	1			

The Assessment System Self-Evaluation

The ASSE, developed by the Pearson Assessment Training Institute, generates a qualitative measure of the status of implementation of various assessment practices that are relevant to the development of a balanced system of assessments. The ASSE is organized into seven action steps and under each action step there are several sub-steps. Understandably, the implementation of any practice is an on-going iterative process and the ASSE is intended to capture the status of implementation at just a moment in time.

To tailor the ASSE more closely to the purposes of this evaluation and to generate comparable data, I asked the administrators to consider the entire district as their point of reference when gauging the current status of implementation. To complete the ASSE, each administrator rated every sub-step on the ASSE on a scale of 1-5 with a rating of 1 corresponding to “getting started,” 3 corresponding to “progressing,” and 5 corresponding to “implemented.” The ratings of 2 and 4 were untitled but clearly represented midpoints. From the pool of 9 potential respondents, 7 administrators returned the ASSE resulting in a 77% rate of participation.

The ASSE generates ordinal data in a numerical format that I aggregated and analyzed using descriptive statistical functions. I report the results of the ASSE at two levels—the level of the individual sub-step and the level of the action step. The results are summarized in Table 9.

Table 11
Status of Implementation from the Assessment System Self-Evaluation

Step	**Description of action step and sub-step**	**Mean**	**SD**
1	**Balance the district’s assessment system to meet all key users’ needs**	**2.31**	
1A	All faculty and staff are aware of differences in assessment purpose across classroom, interim/benchmark, and annual levels, and know how to use each to support and/or verify student learning; that is, to balance formative with summative assessment. We understand what uses can and cannot be made with each level of assessment.	3.00	0.53
1B	Our school board and community understand the concept and need for a balanced assessment system and are supportive of this priority.	2.71	1.16
1C	We have a comprehensive assessment system in place that defines a philosophy of assessment, states the roles assessment can play, and is meeting the information needs of all users. The plan coordinates state-, district-, and building level tests, and supports administrators and teachers in brining assessment balance to the district and its classrooms.	2.57	0.73
	Policies at the district and school levels reflect the value placed on assessment balance and quality, and we have		

1D	identified all of those policies that contribute to balanced and productive assessment, and have a systemic approach to the development and coordination of those policies	2.14	0.83
1E	We have inventoried all assessments used in the district and have categorized them by purpose, standards/targets measured, time of year, etc. for the purpose of understanding the balance we have in our current assessment system.	2.00	0.76
1F	A top assessment priority is to help students develop the capacity to assess their own learning and to use assessment results to help promote further learning.	1.86	0.35
1G	We have an information management system to collect, house, and deliver achievement information to users at classroom, interim/benchmark, and annual assessment levels.	1.86	0.83
2	**Refine achievement standards to reflect clear and appropriate expectations at all levels**	**2.69**	
2A	We continue to refine our local achievement standards, have aligned them with state standards, and have identified our highest-priority learning outcomes.	3.43	1.18
2B	All teachers in the district have received adequate and ongoing support in developing their understanding of the written curricular documents. Teachers are given time to collaboratively plan lessons aimed at accomplishing grade-level/subject expectations.	3.29	0.88
2C	A curriculum implementation plan is in place to ensure consistency in achievement expectations across classrooms. Teachers are held accountable for teaching the written curriculum.	2.83	0.37
2D	Assessment results for all uses are always linked back to local content standards	2.57	0.49
2E	Model/sample lessons and assessments, linked to the content standards, are available and used for professional development.	2.57	0.90
2F	We have deconstructed our standards into knowledge, reasoning, performance skills, and product development learning targets at each grade level for each subject	2.43	0.73
2G	We have transformed the grade-and course-level learning targets that guide classroom assessment and instruction	2.29	0.88

	into student-and-family friendly versions.		
2H	We have verified that each teacher in each classroom is master of the content standards that their students are expected to master. We provide professional support in content areas to teachers when needed.	2.14	0.83
3	**Ensure assessment quality in all contexts to support good decision-making**	**2.75**	
3A	There is a general understanding that quality assessments form the foundation for accurate report card grades and for decision made about students that rely on assessment data.	3.43	0.35
3B	At the classroom level, teachers understand the importance of selecting the appropriate assessment method match to the type(s) of learning target to be assessed in order to help ensure quality results.	2.86	0.35
3C	We have conducted a local evaluation of the quality of all of our assessments, including interim/benchmark and common assessments, if used.	2.57	0.49
3D	We have adopted and can apply the criteria by which we should judge the quality of our assessments, both of and for learning	2.14	0.83
4	**Help learners become assessors by using assessment for learning in the classroom**	**3.05**	
4A	Faculty, staff, policymakers, and community members all understand and embrace the idea of assessment for learning, i.e., student-involved assessment to promote learning.	3.29	0.70
4B	Teachers use assessment information to focus instruction day to day in the classroom and communicate learning expectations to students in language they can understand.	3.14	0.35
4C	Teachers design assessments to help students self-assess and to help them use assessment results as feedback to set goals.	2.71	0.45
5	**Build communication systems to support and report student learning**	**2.74**	

5A	Students are involved in communication about their own progress and achievement status.	3.14	0.99
5B	We have developed standards-based report cards as a means to communicate student progress relative to the targets of instruction, and we provide teachers the support needed to make it work.	3.00	0.86
5C	We understand the value of descriptive feedback used to support learning and know that the best use of evaluative feedback is to judge the level of learning.	2.71	0.45
5D	Teachers know how to offer descriptive feedback to students that will be effective, is delivered during the learning, and is directly linked to the targets of instruction, helping to guide improvement of learning.	2.43	0.49
5E	Teachers understand and apply the principles of sound grading practices, assigning report card grades that are accurate, fair, and are representative of current achievement levels.	2.43	0.73
6	**Motivate students with learning success**	**1.93**	
6A	The classroom assessment practices we use rely on student involvement in assessment during learning to maintain their confidence and motivation.	2.00	0.53
6B	Our faculty, staff, leaders, policymakers, and community understand the power student-involved assessment has to help all students experience the kind of academic success needed to remain motivated, confident, and engaged.	1.86	0.83
7	**Provide the professional development needed to ensure assessment literacy throughout the system**	**3.21**	
7A	Professional development is having its desired impact as our program evaluation shows that we have achieved balance, a high degree of quality assessment, and an increase in student achievement.	3.43	0.49
7B	The development of assessment literacy is offered in a professional development model that allows teachers to learn from each other in collaborative teams and practice in the classrooms as they learn.	3.29	0.70
7C	Our school leaders have developed the assessment literacy they need to maintain the vision, to develop essential infrastructure, and support teacher development in	3.14	0.35

	assessment literacy.		
7D	Leaders are committed to assessment literacy for all. Professional development resources have been allocated to achieve balance in our assessment systems, to have accurate assessments, and to employ assessment for learning practices.	3.00	0.53

Assessment Literacy Inventory

The ALI, developed by Mertler and Campbell (2005), consists of five scenarios that depict assessment practices that classroom teachers typically use. Each scenario is followed by seven questions that correspond to the seven standards of competency as articulated in *The Standards for Teacher Competence in the Educational Assessment of Students* (AFT et al., 1990). To answer the questions, respondents chose one of four possible responses that were presented in a multiple-choice format.

I invited all administrators and all general and special education teachers at the elementary and middle schools to participate. From the pool of 51 potential respondents, 40 participants completed the survey, resulting in a 74% rate of participation. The participants completed the ALI anonymously. In my analysis, I assign letters to the respondents to identify them.

I aggregated all data and analyzed it using descriptive statistical functions. I report the data in terms of accuracy at the level of each participant and the level of the group as both perspectives are relevant. Respondents could potentially obtain a score of 5 correct on each of the competency standards. The mean number correct is 22.88 with a standard deviation of 4.73. Respondents who are significantly divergent from the mean are in bold type. These results are summarized in Table 12.

Table 12
Accuracy of Individual Respondents on Assessment Literacy Inventory

	# Correct on Standard							**Total % correct**	**Standard Deviation from the mean**
Respondent	**1**	**2**	**3**	**4**	**5**	**6**	**7**		
A	4	3	4	5	3	4	5	80	1.08
B	**3**	**2**	**2**	**2**	**1**	**1**	**3**	**40**	**(1.87)**
C	3	3	3	3	3	3	4	63	(0.18)
D	5	2	5	2	4	2	5	71	0.45
E	3	2	4	5	4	4	2	69	0.24
F	2	3	2	3	1	4	4	54	(0.82)
G	3	1	4	2	4	3	3	57	(0.61)
H	4	0	4	3	2	4	4	60	(0.40)
I	4	4	5	3	4	5	4	63	1.29
J	3	2	3	4	3	3	4	83	(0.18)
K	3	2	1	2	2	2	5	49	(1.24)
L	3	1	2	4	2	1	3	46	(1.45)
M	2	3	2	3	2	4	2	54	(0.82)
N	3	5	3	4	3	4	4	74	0.66
O	4	2	3	5	0	3	5	63	(0.18)
P	4	2	4	3	3	4	5	71	0.45
Q	**2**	**0**	**2**	**3**	**5**	**1**	**2**	**43**	**(1.66)**
R	4	3	3	5	2	3	4	69	0.24
S	4	3	5	4	3	3	5	77	0.87
T	5	3	3	5	3	2	4	71	0.45
U	2	2	4	3	3	2	4	57	(0.61)
V	3	2	3	4	2	3	3	57	(0.61)
W	3	4	4	2	3	3	4	66	0.03
X	**2**	**2**	**5**	**1**	**2**	**0**	**1**	**37**	**(2.09)**
Y	3	3	5	1	4	4	4	69	0.24
Z	5	3	2	3	4	3	2	63	(0.18)
AA	5	2	4	4	4	3	5	77	0.87
AB	**4**	**5**	**5**	**5**	**3**	**4**	**5**	**87**	**1.72**
AC	3	2	2	3	2	2	4	51	(1.03)
AD	5	4	4	3	3	4	4	77	0.87
AE	5	1	4	3	5	4	5	77	0.87
AF	4	4	4	5	4	3	5	83	1.29
AG	**5**	**1**	**3**	**3**	**1**	**2**	**3**	**51**	**(1.88)**
AH	3	3	3	3	3	4	2	60	(0.40)
AI	3	2	4	4	3	2	4	63	(0.18)
AJ	3	5	4	4	4	3	5	80	1.08
AK	3	2	3	4	3	3	5	66	0.03
AL	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**100**	**2.56**
AM	3	3	4	3	2	3	4	63	(0.18)
AN	2	3	5	5	3	4	3	71	0.45
% correct on aggregate	**70**	**52**	**71**	**69**	**59**	**61**	**78**	**65**	

Semi-structured Individual Interviews

I conducted semi-structured interviews with all nine administrators. The interview transcripts were coded and analyzed using an open coding in which themes emerge from the data and a constant comparison method in which themes emerge from an on-going comparison of the data with the theoretical literature. I use the overarching constructs of my model of a balanced system of educational assessment to organize the results.

A Balanced System of Educational Assessment

Early on in the interviews I asked each administrator to explain what the concept of a balanced system of assessment meant to them and to identify which elements were most important. Three central themes emerged including: 1) a balanced system of assessment is comprehensive with a variety of assessments to gauge a wide variety of abilities, 2) a balanced system provides useful information in a timely manner, and 3) administrators and staff need to embrace the use of data.

All administrators, to one degree or another, mentioned that they wanted assessments that could capture “a well-rounded sample of a student’s abilities” (BL3). Logically the majority of the assessments would assess academic skills and be “fine-tuned and provide information about specific aspects of learning, such as computational fluency, conceptual understanding and problem solving….the ability to apply mathematical concepts to the real world” (CO2). However, they also wanted to include performance assessments that could capture a student’s “artistic, performance, theatrical or verbal abilities” (BL3).

To get a good picture of a student's abilities, they want assessments to help them understand "process more than product" (CO3). This perspective was especially true when assessing very young children. In those instances they advocated for assessments that incorporate "observation and use of manipulatives…to observe how they perform. Maybe take a picture of what they are going through" (BL3). Another example they gave was running records, an assessment technique that involves listening to a child read and recording the specific errors that he or she makes. One administrator described how staff keep a "series of running records on a child and it is great for progress monitoring for those kids that we are trying appropriate interventions for" (BL4).

Only one administrator expanded the meaning of "abilities" to take into account those students that the educational system has labeled as "disabled."

> A balanced assessment would be a fair assessment for children with disabilities...So, for example, if a student has auditory processing difficulties or processing speed difficulties then an oral assessment is not going to be the best measure for him or her…To be fair it does not have to look the same for every kid. (CO1)

These students do pose unique challenges when it comes to assessing their abilities. To fairly assess them, a system of assessments needs to be very comprehensive in scope.

Their model of a balanced system included "assessments that can give the teacher information that they can use in their planning but also provide the school and district with a bigger set of information in terms of the effectiveness of programs" (BL1). They frequently referred to generic formative and summative assessments as part of their overall system. Most of them mentioned MAP, the interim-benchmark that they have recently incorporated. Overall, the administrators mentioned the evaluative use of assessment much less frequently and some didn't mention it at all.

Another dominant theme to emerge was that assessments had to be useful. For most administrators, the usefulness of an assessment was measured in terms of how many different stakeholders could use the data from an assessment. A "useful" assessment was also "instructional."

> I think good assessment is also instructional. So you think of performance assessment…students are learning through the assessment process…it is not post-mortem…after the fact…what did you learn?...it can also be learning itself. (BL2)

"Useful" also implied that the assessment was "meaningful to them (students) and to teachers as their instructors and administrators to drive instruction and look at where they are and what needs to be tweaked or fixed or added" (CO5). One administrator described her efforts to help young students understand assessment.

> I learned it made sense to explain things to students even if they are younger just so that they know we are doing this to see how you are doing compared to everyone else….to show how good you are in something and we always tried to put it in a positive tone. (CO5)

To help staff make meaning of assessments, the district has committed some of their in-service training time to data summits that bring staff from different schools together to help them understand "why it (assessment) is important and it is not a waste of time….and how it ties into their instruction" (CO5). These conversations continue back at the elementary schools, where staff participate in weekly data-team meetings. These efforts all aim to support a culture where assessment is seen as a useful endeavor.

Embracing the use of data was a third theme that emerged. These administrators universally expressed that they need to find ways to help staff "own the data" (CO3). This administrator made an interesting comparison between the staff overall acceptance of MAP data versus their resistance to MCAS data to illustrate this point.

> Because it (MAP) is not state imposed, we own it a little more. I think we tend to say, “Oh, look at that! That is interesting because it is ours.” But all of the Data Warehouse information that we get about MCAS…. we are not embracing it because we didn’t do it ourselves and we didn’t see it happen and we didn’t generate it. So I think there is that benefit…. which is kind of a silly benefit. (CO3)

These administrators acknowledged that staff are at different places in embracing the use of data and speculate that some resistance is related to “the fear of not understanding it” (CO5). One administrator found that she was more successful in getting staff to embrace data when “the information that they were getting back is very specific to the student ….and specific to certain areas so they know exactly where to zero in” (BL1).

This administrative team sets a model for their staff in their commitment to incorporate data and develop a balanced system of assessment as a key feature of their school system. Unique characteristics of this district make this commitment especially important. For years this district has adopted a model of grouping students heterogeneously by ability at all grade levels. This stance is a key defining aspect of their learning culture. This commitment to heterogeneous grouping will put an added premium on good assessment information. One administrator aptly summed up their situation.

> I think in this environment….a system that is so committed to heterogeneous grouping…. that the use of data and assessment is more critical because you have to meet every child in every classroom at their level to differentiate effectively and absent data that is nearly impossible to do. (CO2)

After exploring the broad concept of a balanced system of educational assessments, I probed deeper with questions targeted at specific aspects of a balanced system. Through this more focused questioning I wanted to understand their perspective on these practices and see what themes emerged. I also wanted to gather information about the current status of implementation of various practices.

Comprehensive Range of Assessment

As I probed into their use of assessment for formative purposes, several themes emerged. These administrators all saw the real benefit of incorporating the formative use of assessments on a frequent basis and wanted to increase its use. Building-level principals were keenly aware of the challenges inherent in incorporating formative assessment. Their reflection on their efforts highlighted the theme that the success of their efforts was closely tied to the willingness of their staff to adopt new practices. All of them acknowledged that they were currently mired down because of old habits. One building-level principal described their current status.

> We are still working on it (formative assessment)…it is a focus and we use that terminology. I am asking teachers to really focus on it on a daily basis. And we meet some times as teams talking about that and go in and watch a lesson talking about that again. Does everybody have it under our belts so that we are really consistent and sure of ourselves? No, but we are working towards that…. I think there is a history of following the (math) book, not necessarily the standards or the curriculum. I think there is a cultural history that is hard to break. (BL3)

Administrators described new initiatives that they were implementing to help their staff adopt new forms of assessment. Last year in one of the elementary schools, the staff piloted a computer-based program, identified only as IXL, which is a website that is designed to reinforce certain math skills. As the child works on specific math problems, the system stores information about the child’s performance over time on certain skills. Teachers can interface with the program and get information about the child’s response to certain types of problems and his or her progress over time. Teachers are beginning to use this information to better target their instruction to the needs of individual students. They reported that they like this program and have advocated for its continued use this year.

By all accounts, staff and the majority of the administrators were receptive to using the district’s newly adopted interim-benchmark system, MAP. There was no consensus as to whether this was rightly categorized as a formative or summative use of assessment. The administrators noted the advantage with MAP is that it can track student progress over several years relative to expected rates of growth and claims to be predictive of achievement levels on MCAS. The disadvantage is that teachers cannot view the mathematical problem or the child’s response and only receive a final score.

Apart from these 2 computer-based systems, there was no mention of other practices that were consistently used throughout the district for formative purposes. Administrators noted that there were individual teachers who were very skilled at incorporating formative practices, such as conferencing with students and using portfolios of work samples. They hoped that these exemplary teachers would serve as role models for others but the district does not have formal process of mentoring in place at this time.

With regard to the summative use of assessment, the administrators again noted the value of these assessments as part of a balanced system of assessment. A central theme that emerged is that they are dependent on external sources for their repertoire of summative assessment practices. They are also in a state of flux due to the demands to transition from the 2004 to the 2011 version of the Massachusetts curriculum frameworks.

All administrators reported that their staff rely on the end-of-unit tests that are available through the math textbook and other materials that were part of their commercial curriculum. A building-level administrator summed up the situation in one of the elementary schools.

> This type of information (summative) teachers would mostly get from the unit…from the tests. So in our upper grades 3-6 they use Scott Foresman and they would be using the chapter test to make those determinations and then in the lower grades …they use *Investigations* and those assessments that are in that. So really we rely on the things that are in place in the curriculum, not necessarily something we are generating. (BL1)

By all accounts, these end-of-units tests are administered at the discretion of the classroom teacher rather than by a schedule that is set by the administration.

Relying on end-of-unit tests is complicated by the very disjointed math programs that are currently in place in this district. In Kindergarten through 2nd grade, this district uses *Investigations*, an instructional program that is geared towards exploratory learning with a focus on developing conceptual understanding rather than by a more traditional sequential skill development approach. At 3rd grade the district shifts to a Scott Foresman series that is characterized by a more traditional approach. In 7th grade there is another shift which fragments the continuum of the math programs even more.

The end-of-unit assessments that accompany each program are quite different from one another in their format. A theme that emerged is that the administrators are frustrated and at a loss to understand if an apparent dip in performance at the grade levels when the curriculum shifts is related to gaps in instruction or just a change in the format of the summative assessments. One administrator captured the essence of the group’s thinking.

> If they (students) have done well in *Investigation*, they get to 3rd grade and realize, “I am not such a good math student!” How awful. And these teachers are saying, “These kids don’t know math”. Is it because they don’t know the systems of Scott Foresman? These kids have learned this body of understanding through *Investigations* and we are not applying it and using it. Do they know math or are they too dependent on the teaching structures? (CO3)

Another theme that emerged is that the shift to the 2011 version of Massachusetts curriculum frameworks is impacting all components of their instructional program and straining their capacity to respond. In the area of assessment, this shift is necessitating an overhaul to their grading system. The scope of this project creates many problems in a district this size and some administrators questioned the role of the state in tackling this.

> We are hoping that the state or a team somewhere across the state will start the report card piece so that we don’t have to start from scratch. But I think everyone in other districts is thinking the same way… “Well, we’ll just see if somebody else does it.” I wonder if the state will come up with something as they have with a lot of other things. (CO5).

Their desire and need to collaborate with educators beyond their district on this and similar projects was a theme that these administrators frequently expressed. A similar problem relates to the coherence of their assessment system to their curriculum. That will be explored in the next section.

These administrators use the results from the MCAS assessments for both summative and evaluative purposes. The majority of them acknowledged that MCAS produces a vast amount of useful data, but “3rd grade is too late to wait to get assessments to know what you need to do” (CO5). As they explained how they analyze and interpret the MCAS, a common theme was that they “lack data traditions” (CO3). In their own words, they have not answered basic questions: What MCAS information gets to teachers? Who is getting that information? Which reports does the administration like? What is our process of reaching conclusions? What do we do based on our conclusions?

There was a consensus that they wanted to do more than just “admire the data” (CO3) and know that they need to build a process by which they consistently use data in ways that translates into informed action.

Coherence Amongst Components

Coherence is typically conceived of as the alignment of the curriculum, instructional practices and assessment components of an educational program to a common set of learning standards. In most school districts, the work needed to align the components is typically spearheaded by curriculum coordinators. In this district there is one administrator who works half-time as the district’s elementary curriculum coordinator and half-time as a technology integration specialist at one of the elementary schools. At the middle and high school level, a department head leads the district’s curriculum efforts in each content area.

From the outset of my study I understood that these administrators had already identified that the components of their math program were not well-aligned and they were taking action to address this gap. Their English Language Arts program had been in a similar state and they devoted their efforts over the past two years to aligning the various components of that program. With that work behind them, they were now turning their efforts to revamping their math program. This year, under the direction of the elementary curriculum coordinator, they formed a Math Curriculum Alignment Study Team (Math CAST) whose goal was “to assess the current status and chart an action plan that will endeavor to incorporate the systematic use of data, a standards-based curriculum, a system of tiered instruction and intervention, and enhanced family and community involvement" (CO3).Through their work, they want to "decide what the indicators (of performance) are at those grade levels so that our four elementary principals will approach it (assessment) in the same way" (CO2).

With that effort underway and the district clearly in a state of transition, I did not probe deeply to evaluate this aspect of their assessment system; however, I did ask each administrator a single question focused on this area. In response, one theme emerged. The shift to the 2011 version of the Massachusetts curriculum frameworks is straining their capacity as administrators. For example, the shift to the 2011 version of the Massachusetts curriculum frameworks is also upturning any work that they had done to align the curriculum, instructional and assessment components of their math program. Although they were not content with their current status of alignment, they had spent resources aligning their curriculum materials and the assessments to the 2004 Massachusetts frameworks by "matching the textbooks to the standards and making it public on the district website" (CO3). The shift to new standards has expanded the work of the Math CAST group and they have to decide if they need or can afford a new set of math curriculum materials for grades pre-kindergarten through 6th grade. They know this will be a multi-year project to achieve some coherence within their math program.

A Robust Capacity for Data Management

A robust capacity for data management implies that a variety of stakeholders can access the result of assessment in a timely manner to inform key decisions. I asked each administrator to respond to the question: To what extent does your current system of data management provide you with timely information to inform decisions? Two central themes emerged from my analysis of their response: 1) technology will play a pivotal role in developing a data management system and 2) developing a data management system is a daunting task that will consume significant resources.

Over the course of several years, this district has incorporated various assessments that have resulted in a patchwork of data management supports. A list of the major assessments and their supporting data management system illustrates the chaotic state of current affairs. MCAS tests results are accessed through the Data Warehouse, a website maintained by the Massachusetts DESE. The results from MAP are accessed through the Northwest Educational Association's website. Fluency data associated with the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), a program for benchmarking and progress monitoring in reading, is accessed through a database managed by the University of Oregon. The district report card system is not straightforward. At the elementary level teachers maintain their own grade books and enter their summative grades into individual Excel spreadsheets that are then printed on a periodic basis. Grades on the report card reference the 2004 curriculum framework standards. At the middle school teachers can enter data into GradeQuick or maintain grade books. Grades on the report card do not reference the standards and appear as just a single score for the content area. A building-level administrator described how the data team meetings work given the current system of data management.

> It is messy right now. At this point I have a three-ring binder that is divided into grade levels and however the teacher brings it (data) to the meeting is how I stuff it into my little book…The problem has been how to translate that into looking at kids longitudinally because we are laying multiple different pieces of paper next to each other…. but we are working on it but that is where we are right now. (BL4)

To make matters worse, teachers cannot access some of these web-based sites. This lack of direct access is frustrating and puts a burden on the building-level administrators and IT support personnel to provide the data. Even teachers who do have access don't all have the requisite skills to maneuver around the sites. One administrator summed up their current status.

> Some of the teachers are able to…they can log in (to the Data Warehouse) but this is where you get… because when you get into the Data Warehouse, they get overwhelmed. They really need the time to learn how to do it and get data. So instead what is happening now is that the principals are grabbing the data and giving it because the principals actually have the time to absorb what is it…. that is why I push for them to have access….I personally think that there is a power that you lose when you don't let the teacher directly connect to the data. (CO4)

These administrators shared that they know that they have "to integrate the pieces" (BL4) and that they have to partner with the instructional technology support personnel to do so. The district's technology department can serve as "the basket that holds the information" (CO4) but it is only the container. The administrative team needs to define the questions that will shape the container. Currently they are at the point of "trying to ask the right questions…and from the questions determining if it is something that can be created" (CO4).

There was universal agreement amongst all administrators that creating a better data management system is a daunting task for this district. They have only one administrator in the capacity of technology support. Even if the district accesses external sites, such as the Data Warehouse, as a core part of their data management system, that will not significantly reduce the burden on local resources. The reality is,

> It [the Data Warehouse] is offered free, but in what sense. It is free to use their servers and to upload the data but the work that goes into it….that is not free….When we start to think about it…if we are going to upload the data, we

> still need a general basket here to dump everything that we are expected to upload over there….we are still going to need that big place here to allow it. (CO4)

The district's administrators engaged in an initiative to resolve this at the time of this study.

High-Quality and Diverse Assessments

The use of high-quality and diverse assessments ensures that all students, including those who have been identified with learning disabilities or are from cultural, linguistic, or racial minorities, can be accurately and fairly assessed. Based on the Massachusetts Department of Education school profile report for the 2010-2011 school year, the student population in this district was 94.8% White, 2.4% Hispanic, and under 1% in all other categories. Less than 1% of students had limited English proficiency, 25% qualified for free or reduced-fee lunch, and 16% qualified for special education services.

In spite of the rather homogenous composition of the student population, the district needs to ensure that all assessments are of high-quality, implying that the assessments are characterized by strong psychometric properties. I asked each administrator, if in their opinion their assessment met this standard. By their own account, their ability to genuinely answer this was beyond the scope of their expertise and they rely on outside entities to address that issue. In the instance of MCAS, this is a state-mandated assessment and it is, therefore, taken for granted that the Massachusetts DESE verifies that this assessment is high-quality. MAP is developed by the Northwest Evaluation Association and the psychometric qualities of this instrument are available through the companies' supporting documents. The district's administrators did not mention that they had looked at that aspect of the assessment when they decided to include it in their repertoire of assessments. The IXL program is a web-based technology from the IXL company located in San Mateo, CA. A review of their website did not reference any psychometric properties relating to this program. The end-of-unit tests that are used by many classroom teachers are part of the curriculum materials and there is no information of this nature included in the teachers' manuals.

Administrators had more to share in regard to how they ensure that they fairly assess the special education population. MCAS is considered a "one size fits all" (CO1) test; however, the district rarely opts to have their special education students take the MCAS-Alternative assessment. The rationale for this decision is rooted in

> the philosophy that if we ever want them to pass it (MCAS), we have to give them the same opportunity as everybody else to practice it every year. We can't give them the MCAS-Alt for all the years and then in 10th grade expect them to be able to do the MCAS. They are already at a disadvantage and they haven't had the opportunity to be through the experience 7 times before. So if we think a child is eventually going to be able to handle the MCAS, then they take the MCAS. (CO1)

Last year they had only one student in the entire district take the MCAS-Alt and this year they have only three students opting for this version of the test.

When it comes to other types of assessments that are frequently used with special education students, there are protective measures in place that ensure the use of high-quality and diverse assessments. The legal requirements for the initial and re-evaluation of special education students mandate the use of tests with adequate psychometric properties. There are also mandates regarding the need for frequent assessment. As noted by one administrator, "If kids are not making progress based on their goals and objectives on the IEP, by law, we have to re-look at them and reconvene the team" (CO1).

Minimum Burden

A well-balanced system of assessment places a minimum burden on students and staff to develop, obtain, analyze, interpret and use assessment information. I asked each administrator to what extent their current system met that standard. There was a general consensus that they did not meet that standard. However, one administrator epitomizes their willingness to engage in the on-going review of current practice that will enable them to be more efficient. Despite almost universal support for the MAP assessment from the administrative team, this one administrator thought it was redundant and wanted them to reconsider their decision. She wanted them "to dig deeper into Data Warehouse and not have our children clog up our computer labs with MAP tests again. Let's be smart adults and not do this to children. Let's use what we have on hand and do a little more with that" (CO3). Her plan was to bring her concern back to the larger administrative team for their consideration.

Overall, these interviews with each administrator were very informative. These administrators took advantage of the opportunity to self-reflect on their strengths and weaknesses. Their commitment to leading their system was also evident throughout the conversations.

Despite my efforts to conduct semi-structured interviews with numerous teachers, only one agreed to be interviewed and one other submitted a written response to the interview questions. This under-representation limits my ability to draw conclusions from the data. However, I summarize the responses of these participants to glean useful information. I identify them as T1 and T2.

Assessments Used by 6th and 7th Grade Teachers

These two teachers report using a variety of assessments, including end-of-unit tests, MCAS and MAP. They expressed an overall favorable impression of the MAP test. They get results in a very timely manner, sometimes within a day or two of testing, and it is easy to access the NWEA web site. The MAP test were characterized as being “less political” (T2) than MCAS.

With regard to report cards, one teacher described the situation this way: “I can have access to previous report cards but find most recent ones highly inaccurate” (T1). The other teacher echoed this feeling and shared that grades on report cards appear to be very subjective. Neither of them reported using them to place students or to facilitate the transition of students.

These two teachers differed in terms of their reported facility with analysis of assessment results. One of them shared.

> The analysis of all this data is pretty much up to me to do on my own time, meaning that there is very little time without students provided for this which I find contradictory given the importance and emphasis that is placed on having “all this data”. I am not a statistician. I wish that someone would look at the data and provide me with an analysis that could tell me what the statistically relevant trends are regarding all this information. For example, if the last couple of years geometry scores are lower than some others does that mean that I did not address this? Is it in the expected range or is it a problem? Is it the class makeup those years? Is it this? Is it that? Sample Size? Is it adequate to determine anything? Is it because I had 15 % of my class with a Math IEP? (T1)

The other reported well-developed skills in the use of data.

Assessment to Facilitate the Transition

Based on the information provided by just these two teachers, MAP results are used to place all in-coming 6th graders, especially students coming in from out-of-state.

Each year the district has moved up the time when students take the fall MAP test and it is now given in the first or second week of school. Reportedly this was done intentionally to maximize the use of this data to make decisions about placing students.

The district has developed a transition sheet that 6th grade teachers fill out on each student. Both teachers questioned the value of this form to facilitate the transition of students to 7th grade. Historically, the 6th and 7th grade teachers have not had the chance to get together; however, these teachers noted that the district is making an effort to use some of the professional development days in the upcoming year to facilitate face-to-face conversations.

In summary, these two teachers reported that they use a variety of assessments. In their opinion, the report cards are subjective and do not provide a valid measure of student performance levels. One teacher reported using MAP results to place students at the appropriate ability level.

CHAPTER 5

DISCUSSION

Introduction

This educational era is characterized by a desire for and expectation that all students attain high levels of academic proficiency. A key feature of many contemporary school reform efforts is the use of educational assessment to reach that goal (Hamilton et al., 2008; Pellegrino & Goldman, 2008; Ryan, 2002). School leaders are faced with the challenge of understanding and employing educational assessment in ways that genuinely enhance student learning. To meet this challenge, school leaders need to conceive of educational assessment as an integrated system that provides a variety of information to many different constituencies in a manner that enables them to make informed decisions (Chappuis et al., 2010; Pellegrino & Goldman, 2008; Rothman, 2010). There is the expectation that school leaders will be actively involved in the development of assessment systems within their school districts (CCSSO, 2008).

The purpose of this research study is to use a utilization-focused evaluation as the strategic plan for school leaders to study their current system and chart a course of action that leads to overall improvement to their system of assessments. Unlike basic research that is undertaken to discover new knowledge or test theories, a utilization-focused evaluation is undertaken to inform decisions, clarify options and support action (Patton, 2008). Three core questions define the focus of this evaluation. In this chapter I integrate the results gleaned from various methods in relation to each research question. I conclude with implications for practice, policy, and future research.

From the outset of this study, the administrators who agreed to participate understood that I would identify strengths, as well as weaknesses, in their current system. Any shortcomings are not to be attributed to a lack of administrative leadership. On the contrary, this district is further ahead than most due to a strong leadership team that is committed to an honest and proactive approach to problem solving.

Summary of Findings

A Balanced System of Educational Assessment

> Question #1: To what extent do we currently have a balanced system of math assessments in pre-kindergarten through 8th grade?

To summarize the findings in relation to this question, I integrate the data and results from the CLMAP, the ASSE, and the semi-structured interviews with administrators. By triangulating these sources I increase the likelihood that the findings are well-synthesized and definitive. I use the over-arching constructs of my model of a balanced system of educational assessments to organize my report of the findings.

I asked each administrator to share their concept or model of a balanced system in order to contrast it with the model of a balanced system of assessments that is the premise of this study. Integrating all of their responses from the interviews and the ASSE, these administrators conceive of a balanced system of assessment as having assessments that are primarily used for formative and summative purposes. The use of assessments for evaluative purposes was mentioned by very few of them. Coherence to a set of standards was not clearly articulated. The need for a robust data management system did not figure prominently. The concept of diversity was mentioned by only one and only in relation to the special education population and no one mentioned the need for high-quality assessments. The concept of a minimum burden was implied by their emphasis on assessments that were “useful.”

The lack of reference to key features of a balanced assessment system does not imply that the administrators will not ultimately want to incorporate them as part of their assessment system. As I probed deeper into each of the dimensions with further questioning later in the interview, they were receptive to incorporating the other key features. It does, however, imply that they need to have a better-articulated vision of the assessment system in order to create a template at the outset that will guide their work.

To some extent, they are aware of this need. Their group rating from the ASSE for statement 1C, which speaks to defining their philosophy of assessment, had a mean of 2.57, indicating that they are midway between “getting started” and “progressing.” Their rating for action step #1 from ASSE, to balance the district’s assessment system to meet all key users’ needs, was rated as 2.31. Only one other action step was rated lower than this one.

It is interesting to note that they articulated constructs that were assumed in my model but not as well-articulated. They included dimensions of how the data would be used. Placing themselves in the role of the consumer, they advocated for an assessment system in which the staff felt that they “own the data.” To own the data implies that the staff see their role in generating the data rather than having it imposed on them. The role of the student as a consumer was not mentioned.

Comprehensive Range of Assessments

A comprehensive range of assessments implies that assessments are used for formative, summative, and evaluative purposes. In a well-balanced system, most of the assessments are used for formative purposes and they are administered frequently. Fewer assessments are used for summative purposes and typically are administered less frequently. Assessments used for evaluative purposes are fewer still and typically require comparing data gathered over a period of months and years.

The Formative Use of Assessments

In the interviews, administrators expressed universal support for the formative use of assessments. This support is echoed in their self-rating of 3.29 on statement 4A from the ASSE, referring to faculty, staff, policymakers, and community members all understanding and embracing the idea of assessment for learning. Their self-rating implies that they are moving beyond “progressing” towards “implemented.”

When it comes to understanding the power of student-involved assessment, a very potent type of formative assessment, their self-rating drops to 1.86 as noted on statement 6B from the ASSE. Student self-assessment is not a top priority as reflected in the rating of 1.86 on statement 1F. They acknowledge that as a district, they have made the least progress on action step #6, to motivate students with learning success, with a rating of 1.93.

There are contradictory measures as to the extent to which teachers are implementing formative practices in their classroom. Based on the interviews, the administrators report a very limited use of formative assessments, essentially by just a few exemplary teachers. They have a few initiatives that they are putting in place, such as IXL, but it is in place in only one elementary school. MAP testing is only at a few grade levels. This level of implementation does not appear to be consistent with the administrators' self-rating on action step #4 from the ASSE for action step #4. The goal of this action step is to help learners become assessors by using assessment for learning in the classroom and the administrators gave themselves the self-rating of 3.05. On Statement 4B, relating to their teachers' use of assessment to focus day-to-day instruction, the self-rating was 3.05. While these rating imply that the district is "progressing", the descriptions of the actual examples in the district imply that a rating closer to the "getting started" end of the scale may be more appropriate.

The CLMAP provides information on the use of formative assessment from the teachers' perspective. Overall, the CLMAP indicates that the teachers do not have a solid grasp of what characterizes the formative use of assessment. The teachers reported that the most commonly used formative assessments were tests from commercial curriculum material and quizzes. They also reported these assessments as some of the most commonly used summative assessments. These assessments are typically considered to be used more appropriately for summative rather than formative purposes. They identified reviewing daily work as the most useful practice. Other formative assessment practices, such as student conferencing, were reported by only one teacher.

These findings reflect the challenges inherent in implementing the formative use of assessment. Despite their desire, there are some significant hurdles as they move towards implementing formative assessment as a component of a balanced system of assessment. There are gaps in the staff grasp of the appropriate means to assess students in a formative manner. This leads to a reliance on assessments, such as interim-benchmark assessments, for which there is no research to support the claims of a positive effect on student performance.

The Summative Use of Assessment

The summative use of assessments featured prominently in the reports of all participants from all sources. In the interviews, administrators noted that their staff rely on the end-of-unit test and quizzes as their primary source of classroom-level summative assessments. The teachers' self-report on the CLMAP, confirmed this and underscored that they perceive these assessments to be generally helpful.

Administrators were more open in questioning how helpful the end-of-unit tests really are. Administrators identified problems associated with relying on these assessments, including that they were minimally aligned to the 2004 version of the curriculum frameworks and there is no work at the time of the study to align them to the 2011 version of the curriculum frameworks. The self-rating of 2.57 for statement 2D on the ASSE, reflect their awareness that they are only approaching the level of "progressing" on the task of always linking their assessment results back to local content standards.

Teachers reported on the CLMAP that they use teacher-generated tests and quizzes. This is to be expected, however, it raises some concerns. The results from the ALI provide a measure of the overall level of competency in regard to the skill of developing assessments for instructional decisions. The aggregate level of accuracy of staff on this standard was 52% and ranked lowest of all the standards and implies a relative weakness in this skill area.

The teachers who participated in the semi-structured interviews reported concerns with the district's grading practices. Their comments cannot be taken as representative of the rest of the faculty given the very limited number of participants. Nevertheless, their comments need to be taken into account. They both noted that they cannot trust that grades on report cards are a valid indicator of student performance. They attribute this to a subjective, rather than an objective, approach to grading students. The report card itself is not an issue.

The teachers who completed the CLMAP identified MCAS as a summative assessment that they use almost as commonly as end-of-unit tests. They also gave it a generally favorable rating in terms of usefulness. From the administrative perspective, MCAS was also a core summative assessment; however, their perceived value of this assessment was affected by two factors.

The first factor is beyond their control but is important to note. They expressed the frustration that MCAS begins in 3rd grade and this postpones the opportunity to use the results to guide interventions in the earlier grades when they can be more effective. This implies that the administrators rely on MCAS to inform critical decisions.

The other factor, which they can affect, is related to their district's "lack of data traditions." As an administrative team, they have not identified how they want to consistently use MCAS reports. They acknowledged that they need to decide how they want to use the Data Warehouse as a resource for analysis of results. They also acknowledged that they need to simply find the time to sit together and review the results.

The findings in regard to their current use of summative assessments highlight some areas where these administrators can focus their efforts. There is the need to establish a more consistent method for using the summative assessments in their repertoire. Currently the end-of-unit tests are administered at the teacher’s discretion. An established schedule would facilitate a comparison of results across classrooms and schools. A consistent method for analyzing MCAS results will enhance their chances of using the results to inform decisions. Professional development aimed at improving the staff level of competency for designing tests is warranted. With better-developed skill levels, they will be less dependent on sources external to the district for these tests.

The Evaluative Use of Assessment

Currently, the evaluative use of assessment in this district appears limited. Based on the interviews with the administrators and from results of the CLMAP, MCAS is their primary assessment tool for this type of assessment. Teachers reported teacher-made tests in this category but this is more likely a reflection of their misunderstanding than practice. The ASSE did not solicit information on the evaluative use of assessment. This is a significant oversight in the design of this survey.

The limited use of assessment for evaluative purposes at the district-level underscores the vulnerability of the school district. In this era of accountability there is an increasing use of these measures to evaluate districts from afar. With no means of providing their own measures to gauge the effectiveness of programs or personnel, school districts are in the position of defending what they know to be effective programs in the face of external measures that imply otherwise. To complicate the issue, the task of developing local measures often exceeds the expertise of staff. Developing better practices in the evaluative use of assessment will require new efforts at all levels of the educational system.

Coherence

Throughout the interviews these administrators referenced Math CAST, their district's initiative to revise their math program. These efforts are rooted in their understanding of the importance of the coherence of the curriculum, instruction, and assessment to a common set of learning standards. Their self-rating of 2.69 on Action step #2, the action to refine achievement standards to reflect clear and appropriate expectations at all levels, reflects their awareness that they are not yet at the level of "progressing" in their efforts to align the components of their math program. This action step ranks fifth out of the field of seven steps. These administrators were well-aware that a project of this magnitude requires a long-term commitment and had factored that into their plan.

An analysis of the ratings for the individual statements within this category highlights areas of relative strength and weakness. Their rating of 3.43 on statement 2A implies that they are furthest along in refining local achievement standards and in identifying their highest-priority learning outcomes. The rating of 3.29 on statement 2B implies that they are also relatively further along on the task of supporting their teachers in understanding the written curricular documents. They are not as far along in verifying that each teacher is master of the content standards as reflected in their rating of 2.14 on statement 2H.

The shift to the 2011 version of the Massachusetts curriculum frameworks from the 2004 version is affecting all school districts. On the one hand, the shift is timely in the sense that this district is already in the process of revamping their math program. On the other hand, the shift affects other content areas and will likely necessitate that they revisit the work they have already done in the area of English Language Arts. The administrators did not elaborate on how they plan on meshing these projects.

These findings highlight how projects of this magnitude stretch, and sometimes exceed, the capacity of small school districts and lead to some vexing questions. Is it realistic to rely on only one half-time elementary curriculum coordinator to spearhead the efforts to align their entire elementary curricular program? Do they need to dedicate more district staff to these efforts? Are there ways to collaborate with other districts that they haven’t already tapped to work on this together? What is the role that the DESE or the regional assistance center should play in supporting this work at the level of the individual school districts? Would it help solve these problems if the state’s efforts to regionalize these small districts into larger collaborative groups came to fruition?

A Robust Capacity for Data Management

Throughout their interviews these administrators painted a picture of a data management system with some significant short-comings. Many of their assessments are supported by their own web-based system and the sites are linked with one another. The teachers do not have direct access to some of the sites and have to rely on building principals to give them assessment results. Some staff members reportedly lack the computer skills to navigate to and around some sites. Their rating of 1.86 for statement

1G from the ASSE implies that they are just past the threshold of "getting started" on developing an information system to collect, house, and deliver achievement information to all users.

The administrators report that the task of increasing the capacity of their data management system is daunting. There is only one administrator charged with maintaining all aspects of the district's data infrastructure and by her own account, is consumed by other projects. In the opinion of this individual, the resources that Massachusetts already provides to districts through the Data Warehouse are helpful to a point, but will not alleviate the need for the district to develop more capacity at the local level.

Unlike the response to many of the other features of their assessment system that need revamping, there is no initiative in place at the time of this study to help move the district towards a more robust capacity for data management. The reasons behind this apparent inaction are unclear based on the data. This may indicate that this feature was not thoroughly assessed with the methods of this study. It may also imply that the administrators are genuinely unclear at this time as to how to proceed.

High-Quality and Diverse Assessments

Due to the homogeneous composition of the student body, the district has fewer hurdles on their path to develop a balanced assessment system that satisfies this criterion. Their primary source of diversity is their special education population and they report having practices in place to fairly assess these students. It is interesting to note that not one administrator brought up the issue of diversity in relation to the school choice students from out-of-state, although it is safe to assume that these students present differently in some aspects from their in-district students. For instance, it is likely that Massachusetts and the sending state have similar, but not identical, curriculum frameworks and formats for their state-mandated assessments.

On the issue of high-quality assessments, they put their trust in the test developers to ensure the quality of the majority of their assessments. In the case of MCAS, this trust has some solid foundation. In the case of other assessments, it is more difficult to gauge the quality because it is not always reported, as in the case of end-of-unit tests. According to the ASSE, they have not adopted and applied criteria to judge the quality of either formative or summative assessments as reflected in their rating of 2.14 for statement 3D. They do, however, have a general understanding of the need for high-quality assessments as reflected in their rating of 3.43 on statement 3A.

It appears that they have not focused their efforts on this feature of an assessment system at this time and have no plans to do so in the near future. It simply is not a priority. It was not stated as such, but perhaps this stance is related to the homogeneous composition of their student body.

Minimum Burden

There is consensus on the part of the administrative team in regard to whether they meet the criteria for an assessment system that places a minimum burden on students and staff. Simply stated, they do not. They acknowledged, by their rating of 2.00 on statement 1E from the ASSE, that they are midway between “getting started” and “progressing” in their work towards inventorying all assessments by purpose, standard, and time of year. Developing an inventory is a likely place to start as they focus their efforts on improving this feature of their assessment system.

In summary, as each administrator articulated his or her concept, there was some overlap with the model of a balanced assessment system that is the premise of this study. As a group, they under-emphasized the concept of high quality and diverse assessments. They highlighted other features that they wanted, such as fostering a sense that staff at all levels “own the data.”

With regard to having a comprehensive range of assessments, they currently have a system that is weighted more heavily on the use of assessment for summative purposes than either on formative or evaluative uses. To some extent that is typical of most systems; however in this district, the balance illuminates some gaps in current practice. On the whole, staff members are not well-informed about the nature and use of formative assessments. With regard to the summative use of assessments, they rely heavily on tests that are from outside agencies, such as the commercial curriculum or the state-mandated system. It is important to underscore that the staff who participated reported that they cannot rely on the report card as a valid measure of current performance because they suspect that teachers engage in subjective grading practices. With regard to the evaluative use of assessment, they use MCAS results for this but currently lack a consistent set of “data traditions” to help them identify what results they want to analyze.

They are making a concerted effort to have coherence amongst the various components of the math program through the work of the Math CAST group. By all accounts, this group is just getting underway and plans on working on this project over the course of several years. Their data management system is not well-integrated and staff have some difficulty accessing the systems that they do have. The task of creating a new system is daunting and raises the issue of whether or not the district has the capacity to undertake this. Due to the reliance on externally produced assessments and the homogeneous composition of the student body, the feature of high-quality and diverse assessments is not forefront in their current understanding of assessment. They cannot determine if they have a system that imposes a minimum burden on students and staff because they acknowledge that they do not have an inventory in place at this time.

Assessment to Facilitate Transition

> To what extent are our 6th and 7th grade teachers using math assessment to facilitate the transition of continuing and in-coming students into 7th grade?

As stated at the outset, gathering enough data to adequately answer this question was tenuous given the small pool of potential participants. As it turned out, only two teachers opted to participate and only one could be interviewed. I appreciate the contribution of these two staff members. I am, however, unable to extract valid findings from this limited data. I do attempt to express the issues that they raised.

Both teachers questioned the value of report cards as a gauge of student performance levels. The underlying problem was not with the report card itself but with what they perceived to be subjective grading practices. Results from the ALI substantiate this concern. Standard 5 on the ALI relates to the skill of developing valid grading procedures which use pupil assessments. The aggregate level of accuracy on this standard was 59% correct. The group performed at lower levels of accuracy on only one other standard. Results from the ASSE also reflect there are some gaps in the district’s use of sound grading practices. Statement 5E references the skills of understanding and applying the principles of sound grading practices, assigning report card grades that are accurate, fair, and are representative of current achievement levels. The administrators rated their current status as 2.43. This implies that the district is not even midway in the process of having well-implemented grading practices throughout the district. Due to their concern, neither reported using report cards to facilitate the transition of students into their classroom.

The lack of confidence in grading practices has other implications. Parents rely on report cards as one of their main sources of information about the progress of their children. If sound grading practices are not well-implemented, parents may have an unfounded impression of their children’s performance levels. Given the need for parents to have accurate information and the time that teachers spend filling out report cards, the district may want to focus some professional development in this area to increase skill levels of staff.

These two teachers reported using MAP results as their assessment-of-choice to facilitate the transition of students into 7th grade. Results from the CLMAP indicate that MAP is perceived to be useful by the staff who reported using it as a summative assessment. The majority of administrators also have a favorable opinion of MAP; however, one administrator did question whether or not this test did bring an added value to what they already had and advocated for the district expanding their use of MCAS.

An additional concern in regard to the use of MAP was not raised by either administrators or staff but is found in the research literature. At this time there is no research to support the claim that the use of interim-benchmark assessments, such as MAP, has a positive effect on student achievement. It is possible that MAP may appeal to staff for several reasons. It is easy to administer and only requires students accessing a computer. The results arrive quickly and are easy to access. This, however, does not necessarily translate into an effective assessment that enhances student performance. Based on this concern, the districts' on-going commitment to MAP as part of their repertoire warrants further study by the administration.

Competency in the Educational Assessment of Students

> What is the level of competency of our staff relative to established standards of competency for the educational assessment of students?

A joint task force of the AFT, NCME, and NEA articulated *The Standards for Teacher Competence in the Educational Assessment of Students*. Although dated, they remain the primary point of reference. They state that teachers should be skilled in:

> (1) choosing assessment methods appropriate for instructional decisions
> (2) developing assessment methods appropriate for instructional decisions
> (3) administering, scoring and interpreting the results of both externally produced and teacher produced assessment methods
> (4) using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement
> (5) developing valid pupil grading procedures which use pupil assessments
> (6) communicating assessment results to students, parents, other lay audiences, and other educators
> (7) recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information. (AFT et al., 1990)

The ALI is the primary source of data I analyze in response to this question but relevant portions of the semi-structured interviews and the ASSE also inform the findings. I analyze the results at the individual and group level with reference to the accuracy of response. I organize my findings using the framework of the seven standards of competency but I rearrange their order to rank them from highest to lowest in terms of aggregate level of accuracy.

Standard 7: Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

The aggregate level of accuracy on this standard was 78% correct. Thirteen out of the forty participants obtained perfect scores of five correct answers. At the other end of the spectrum, five participants had scores of just one or two correct answers. Overall, these results imply that staff manifest a relative strength on this standard relative to the other standards.

For purposes of my research, I administered the ALI anonymously to maximize the rate of participation. In the future the district should consider gathering the data in such a way as to identify specific individuals. One advantage is that they could target professional development to specific individuals or tap the expertise of individuals, such as respondents, A, D, and K, as an internal resource. Equally important, they could identify respondents M, Q, and X who had low rates of accuracy. When it comes to issues of unethical, illegal and otherwise inappropriate methods and use of assessment, a low score for even one or two individuals can result in negative consequences for the entire district.

Standard 3: Teachers should be skilled in administering, scoring and interpreting the results of both externally produced and teacher produced assessment methods

On this standard the group had an aggregate score of 71% correct. Seven individuals had perfect scores and eight individuals had scores of one or two correct answers. These results imply that this is also an area of relative strength for the staff. On more qualitative measures, such as the interviews, a different perspective emerges. One teacher commented at length about the lack of ability to interpret data and the time it takes to do this type of work. An administrator expressed the sentiment that overall the district needed a lot of help to properly interpret assessment results.

These differences may reflect the frustrations that these individuals experience when confronted with a set of data rather than a genuine lack of skill. It can be time-consuming to sift through numbers to extract meaning. Some teachers may perceive these more objective measures as a challenge to their own intuitive sense about students honed over the course of their career.

Standard 1: Teachers should be skilled in choosing assessment methods appropriate for instructional decisions

The aggregate score on standard one is 70% correct. An equal number of six individuals had perfect scores or scores of just one or two correct answers. Statement 3B from the ASSE provides an additional measure of this skill from the perspective of the administration. The rating of 2.86 in reference to the staff understanding the importance of selecting the appropriate assessment method to match the learning target, implies a somewhat weaker skill level. The more objective results from the ALI may indicate that the staff are genuinely further along in being capable of making these choices than the administration perceives.

Standard 4: Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement

The skill of using assessment results to inform decisions about teaching goes to the heart of why assessment can be so beneficial. To be unskilled in this regard is significant because it implies that even with robust data at hand, some teachers may not be able to independently translate it into informed action. The group score on this standard was 69% correct. Overall this is still a relative strength. Nine individuals had perfect scores; however, seven individuals had scores of just one or two correct answers. The district is justified in the concern that students in the classroom of participant B, F, and K may not be benefiting from all of the assessments the districts uses. Mentoring less-skilled teachers by their more-skilled peers may be a worthwhile strategy to pursue to help all staff develop competency on this standard.

Standard 6: Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators

The aggregate score on this standard was 61% correct. Only two individuals had perfect scores and eleven individuals had scores of just two, one, or zero correct answers. The higher incidence of inaccurate individuals on this standard may be a reflection of the interplay of skills. Competency in communicating assessment results is dependent on having competency in other skills, such as appropriately interpreting results. If individuals do not have the skills to interpret the data, they most likely will not want or know how to interpret them. To address the shortcomings of the staff on this standard, administrators may need to pair training with this and other standards at the same time.

Standard 5: Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments

The issues surrounding grading practices have already been discussed at some length. The aggregate score of 59% correct underscores the concerns that have been raised. There is only one participant who obtained a perfect score and twelve who obtained a score of two, one, or zero correct answers. The administrative team may need to bring in resources from outside the district to help their staff develop skills in this domain.

Standard 2: Teachers should be skilled in developing assessment methods appropriate for instructional decisions

With an aggregate score of 52% correct, staff manifest a relative weakness on this standard. Four individuals obtained perfect scores while nineteen individuals had scores of two, one, or zero. These results highlight the significant gap in the overall skill level of staff in this domain. Not having the skills to develop assessment methods has implications. If they are developing assessments on their own, as indicated on the CLMAP, those assessments may not be of high-quality. If they are not developing assessments at all because they lack the confidence to do so, then they may have to over-rely on sources such as the end-of-unit tests from the curricular materials. The administrative team is well-advised to make the remediation of this skill area a priority.

When the ALI is taken in its entirety, the group had an aggregate score of 65% correct. Five participants obtained scores that fell in the upper ranges of accuracy, between 80-100%. Twenty-three participants fell in the 60-79% range. Ten respondents were in the 40-59% accuracy and only two respondents were in the 20-39% range of accuracy.

The ASSE provides a more subjective measure of overall assessment literacy of staff and highlights the efforts of the administration to foster assessment literacy through a concerted program of professional development. Action step #7, to provide the professional development to ensure assessment literacy throughout the system, was rated as 3.21 with the implication that this district was further along in this area than in all other areas assessed by the ASSE. The rating of 3.29 on statement 7B reflects the perspective of administrators that they are beyond “progressing” and on their way towards “implemented” in regard to developing the assessment literacy of their staff through a model of professional development that affords staff the opportunity to work in collaborative teams.

In summary, the results of the ALI and other measures indicate that overall there are relative strengths and weaknesses in the current levels of staff competency in relation to the established standards. The majority of the staff are most competent in recognizing unethical practices; however, there are some individuals with significant gaps in this area. The majority of the staff are least competent in developing assessments that are appropriate for instructional purposes. These administrators have devoted resources to fostering the assessment literacy of their staff through professional development and plan on continuing to do so in the future.

Implications for Practice, Policy, and Research

The purpose of this study was to explore the ways in which school administrators can create a balanced system of educational assessments in their school districts. This study was conducted with a team of administrators in one small regional school district with the expectation that it would yield actionable findings. The recommendations integrate the findings in light of the new assessment initiatives that are underway in Massachusetts at the time of this study.

All school districts need to explore ways to harness the power of assessment to enhance student learning. The importance of using assessment effectively is heightened in this district because of its commitment to heterogeneous grouping of students by ability. While the team of administrators appears to have a shared philosophy with regard to heterogeneous grouping, it is less clear if they have a shared philosophy as to the role of assessment. These administrators may consider spending time on articulating their vision for assessment. This process can help them to define the “data traditions” that they want to create and answer the questions that they posed to themselves: What information do they want from MCAS or any other assessment? How will they consistently use that data to inform specific decisions? Who is responsible for providing this data? To create a better balance amongst the types of assessments, the shortcomings of staff with regard to understanding and implementing formative assessment will have to be addressed. Currently staff appear to have some basic confusions about the types of assessment that are appropriate for this use. By all accounts, there are very few teachers who are incorporating any type of formative assessment into their classrooms at this time. A plan to remediate will likely need to include targeted professional development and faculty study groups. The district may consider using resources, such as *Seven Strategies of Assessment for Learning* by Jane Chappuis, as a foundation for this work.

The staff in this study reported their lack of confidence in report cards as a useful measure of the current level of student performance. The district may want to delve deeper into this to understand the extent to which the rest of the staff have this perspective. If this is more pervasive, then the district should consider focusing on developing the competency of their staff relative to grading practices. They are currently exploring a revision to their report card but their work in this area may need to begin with appropriate ways to grade students.

With the Math CAST group, the district has a plan in place to address the lack of coherence amongst the components of their math program. Their work is timely in the sense that it can interface with the transition to the 2011 version of the Massachusetts

Frameworks. Their work is hampered in the sense that the implementation of the Common Core Standards and Assessments is still in the design stage. The district is advised to carefully track the progression of that work of PARCC and the MA DESE. PARCC has an Implementation Workbook to guide districts however only 7 out of the 12 chapters were written at the time of this study.

The work that is needed to address the needs in their data management system is significant and illustrates the lack of capacity that many districts experience. As part of the transition to the new Massachusetts state-mandated assessment system, PARCC is developing a new data management component. At the time of this study, the design of that the data management component has not been articulated. The district may want to wait to undertake any work in this area until the requirements and format of the new system are better-defined.

The district should also consider taking an inventory of all of their current assessments. The inventory would gather basic information on each assessment, including when and to whom they are given and how are the results used. This inventory can then be the basis to determine if there are redundancies and gaps in their current system. An inventory of this nature would be very helpful in the transition to the new state assessment system in 2014. By all accounts, the new state system will incorporate assessments at younger grade levels and types of non-summative assessment at upper grade levels. With this inventory in place, the team of administrators can fruitfully discuss if assessments, such as MAP, do bring an added value that justifies the resources spent on it.

Gaps in the current levels of assessment literacy are a hurdle in this district that can have a negative impact on any of the other initiatives that the district undertakes in the area of assessments. This district already has the culture of fostering the development of their staff through targeted professional development. These administrators may want to consider sharing the results from this administration of the ALI or re-administering the ALI or a different instrument. The purpose of this is to underscore the urgent need for all staff to improve their skills in this area. The district should prioritize their training relative to the gaps, but a logical focus at this time is to foster the ability of staff to design assessments for instructional purposes. This work would complement the efforts focused on improving the use of formative assessments.

There are also implications for policy on a broader scale. Based on the findings from this study, a limiting factor that can impede the successful implementation of the new state-mandated assessment system in Massachusetts is the capacity of a school district to respond. The demands to realign the components of their program or develop a more robust data management system can easily exceed what a small team of administrators can do. The state will need to have a significant role and provide assistance in tangible ways to districts throughout the transition. This is an opportunity to explore how these small districts can be organized into larger collaborative units.

In Massachusetts the transition to PARCC presents new opportunities to consider how the large-scale summative assessment can be integrated into a model of a balanced system of assessment. With the inclusion of non-summative measures and assessments for students in grades Kindergarten-grade 2, it is possible that new assessment practices may become common practice in classrooms. New assessments may employ new technologies that may also effect assessment practices.

There is the need for a new set of standards of competency in the educational assessment of students that incorporate new forms of assessments, such as student self-assessment. With new standards there is the need for new ways of measuring levels of competency of pre-service and practicing staff. The results of this study also imply that there is the need for effective training programs for staff in this domain. The training needs to start at the pre-service level and competence in this area should be part of mandatory teacher competency tests.

In regard to research, there is the need to study the developmental trajectory of the acquisition of academic skills in order to understand expected rate of progress. Rates of progress will need to be fine-tuned for different populations to take into account factors, such as gender or socio-economic status. In the case of the evaluative use of assessment, the interpretation of the results often rest on assumptions about the rate of student progress. If these assumptions are unfounded, then the demands put on school system to meet the expected rate are unrealistic.

Conclusion

It is likely that the reliance on the use of educational assessment as a key component of school reform efforts will continue. With this reliance comes the need to genuinely understand how best to harness the power of assessment to enhance student learning. School leaders play a pivotal role in these efforts. An area of study that merits their attention is to understand the effects of a balanced system of assessment on learning;

however, these efforts are hampered by the reality that a balanced system is not a common practice.

A contribution of this study is that it provides a framework for leaders to follow as they undertake the task of developing a balanced system. The framework of a utilization-focused evaluation proved to be particularly valuable because it enabled this team of administrators to tailor the evaluation to their specific interests and unique characteristics of their school district. The emphasis on utility increased the likelihood that time spent on the analysis would lead to action.

Assessment literacy also lies at the core of any effort to harness the power of assessment to enhance student learning. A contribution of this study is that it provides a measure of the current level of competency for a group of practicing educators in the educational assessment of students. The findings highlight the critical need to enhance the ability of all staff to appropriately administer, score, interpret, communicate, and use assessment results.

There are school administrators working in other school districts who are deeply committed to the success of each and every student under their care. Our understanding of how best to use educational assessment is furthered through the exchange of ideas. The ultimate goal of this study is to support the initiative of other school administrators to solve similar challenges within their school districts by sharing the efforts of this one group of dedicated administrators.

APPENDIX A

CONSENT FORM FOR INTERVIEW

EVALUATION OF THE PRE-K-8TH GRADE MATH ASSESSMENT PRACTICES
IN ONE MASSACHUSETTS REGIONAL SCHOOL DISTRICT

CONSENT FOR VOLUNTARY PARTICIPATION

Rita Detweiler, M.Ed. is the principal investigator and is conducting this study as part of her doctoral work at the University of Massachusetts. The faculty member supervising the research is Rebecca H. Woodland, Ph.D.

Purpose

This study is being conducted as part of Rita Detweiler's doctoral dissertation. Data will be collected that can be used to improve the functioning of the system of math assessments in grades Pre-Kindergarten through 8th grade of my school and regional school district. I understand that the results from this study will be included in Rita Detweiler's dissertation and may be included in manuscripts submitted to professional journals or publications. I understand that the results will not be used in any type of evaluation of personnel.

Voluntary Participation

I volunteer to participate in the interview that is part of a study. I am free to participate or not participate without prejudice. I may withdraw from the study at any time.

Right to Privacy

I have the right to privacy and my name will not be used, nor will I be identified personally, at any time. I understand that it may be necessary to identify participants in the dissertation by position. The small number of participants increases the risk that I may be identified as a participant in the study or the district may be identified.

Interview Process

I understand that I will be interviewed in individual settings with the principal investigator. The questions I will be answering will solicit my views on the math assessment practices within the school district and my role as an administrator or classroom teacher. I understand that these interviews will be audio-taped and later transcribed by someone other than the principal investigator. This individual will have no knowledge of the school district or participants. After the audio tapes are transcribed, the principal investigator will erase the tapes to ensure that no person may be identified by voice. Written transcription material will be maintained and held by the principal investigator for a period of one year after which the material will be appropriately destroyed.

Anticipated Benefit
Results of this evaluation should prove to be useful to the participants in their professional capacity. Concrete recommendations will be made that can be used to guide organizational change.

Contact Information
If for any reason you need to contact either Rita Detweiler, principal investigator, or Rebecca Woodland Ph.D., the supervising faculty member, the contact information is as follows:

Rita Detweiler
200 Lower Rd.
Deerfield, MA 01342
413-687-1750
rjdetweiler@acad.umass.edu

Rebecca Woodland, Ph.D.
Associate Professor, Department of Educational Policy, Research and Administration
University of Massachusetts
111 Thatcher Way
259 Hills House South
Amherst, MA 01003
413-545-1751
rebecca.woodland@educ.umass.edu

____
Researcher’s Name

____
Date

____
Participant’s Name

____
Date

APPENDIX B

CONSENT FORM FOR ASSESSMENT LITERACY INVENTORY

EVALUATION OF THE PRE-K-8TH GRADE MATH ASSESSMENT PRACTICES IN ONE MASSACHUSETTS REGIONAL SCHOOL DISTRICT

CONSENT FOR VOLUNTARY PARTICIPATION

Rita Detweiler, M.Ed. is the principal investigator and is conducting this study as part of her doctoral work at the University of Massachusetts. The faculty member supervising the research is Rebecca H. Woodland, Ph.D.

Purpose
This study is being conducted as part of Rita Detweiler's doctoral dissertation. Data will be collected that can be used to improve the functioning of the system of math assessments in grades Pre-Kindergarten through 8th grade of my school and regional school district. I understand that the results from this study will be included in Rita Detweiler's dissertation and may be included in manuscripts submitted to professional journals or publications. I understand that the results will not be used in any type of evaluation of personnel.

Voluntary Participation
I volunteer to participate by completing the Assessment Literacy Inventory. I am free to participate or not participate without prejudice. I may withdraw at any time.

Right to Privacy
I have the right to privacy and my name will not be used, nor will I be identified personally, at any time. I understand that the surveys are administered anonymously and therefore I cannot be identified in any way.

Survey Process
The Assessment Literacy Inventory will be administered in a group setting. I will have the opportunity to ask questions about the purpose of the study and the survey instrument of the principal investigator in the group setting.

Anticipated Benefit
Results of this evaluation should prove to be useful to the district as they plan for professional development of staff.

Contact Information
If for any reason you need to contact either Rita Detweiler, principal investigator, or Rebecca Woodland Ph.D., the supervising faculty member, the contact information is as follows:

Rita Detweiler
200 Lower Rd.
Deerfield, MA 01342
413-687-1750
rjdetweiler@acad.umass.edu

Rebecca Woodland, Ph.D.
Associate Professor, Department of Educational Policy, Research and Administration
University of Massachusetts
111 Thatcher Way
259 Hills House South
Amherst, MA 01003
413-545-1751
rebecca.woodland@educ.umass.edu

Researcher's Name	Participant's Name
Date	Date

APPENDIX C

INTERVIEW QUESTIONS FOR ADMINISTRATORS FOR QUESTION #1

1. What is your current role in this district and how many years have you been working in this capacity?

2. What does a balanced system of assessments mean to you?

3. What do you consider to be the most important element of a balanced system of assessment?

4. In your opinion, does your current system of math assessments give you and your teachers enough data?
   - To inform instruction on a day-to-day basis
   - To gauge student progress over time
   - To document that students have reached a certain level of mastery and can progress to the next stage
   - To evaluate the effectiveness of curricular programs or personnel

5. In your opinion, are your math assessment aligned with other components, such as curriculum and instruction? Are the components aligned with a set of learning standards?

6. To what extent does your current system of data management provide you with timely information to inform decision?

7. Do you consider your math assessment to be sufficiently high-quality and diverse?

8. In your opinion, does your current system of assessments place a minimum burden on students and staff to develop, obtain, analyze, interpret and use assessment information?

9. If you could change one aspect of your math assessment system, what would that be?

10. Are there math assessment practices that you have tried but have discontinued? If so, why?

11. Do you think there are hurdles or problems that stand in the way of district bringing about any desired change to the current math assessment system?

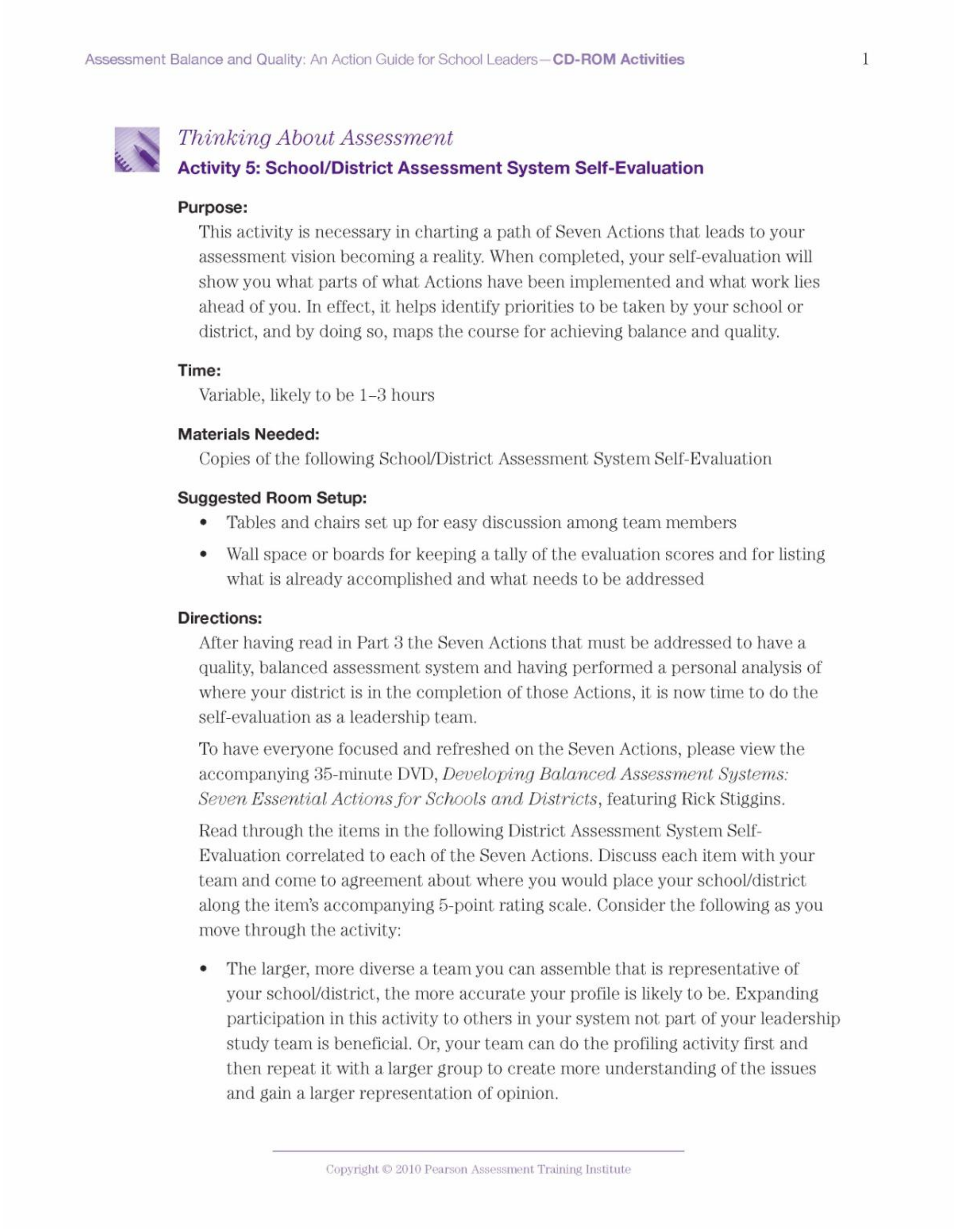
12. Where do you get your information about new math assessment practices?

13. What do you think your district can do to foster an assessment literate culture?

14. Is there anything that we should have talked about but didn’t?

APPENDIX D

ASSESSMENT SYSTEM SELF-EVALUATION



Thinking About Assessment

Activity 5: School/District Assessment System Self-Evaluation

Purpose:

This activity is necessary in charting a path of Seven Actions that leads to your assessment vision becoming a reality. When completed, your self-evaluation will show you what parts of what Actions have been implemented and what work lies ahead of you. In effect, it helps identify priorities to be taken by your school or district, and by doing so, maps the course for achieving balance and quality.

Time:

Variable, likely to be 1–3 hours

Materials Needed:

Copies of the following School/District Assessment System Self-Evaluation

Suggested Room Setup:

- Tables and chairs set up for easy discussion among team members
- Wall space or boards for keeping a tally of the evaluation scores and for listing what is already accomplished and what needs to be addressed

Directions:

After having read in Part 3 the Seven Actions that must be addressed to have a quality, balanced assessment system and having performed a personal analysis of where your district is in the completion of those Actions, it is now time to do the self-evaluation as a leadership team.

To have everyone focused and refreshed on the Seven Actions, please view the accompanying 35-minute DVD, *Developing Balanced Assessment Systems: Seven Essential Actions for Schools and Districts*, featuring Rick Stiggins.

Read through the items in the following District Assessment System Self-Evaluation correlated to each of the Seven Actions. Discuss each item with your team and come to agreement about where you would place your school/district along the item's accompanying 5-point rating scale. Consider the following as you move through the activity:

- The larger, more diverse a team you can assemble that is representative of your school/district, the more accurate your profile is likely to be. Expanding participation in this activity to others in your system not part of your leadership study team is beneficial. Or, your team can do the profiling activity first and then repeat it with a larger group to create more understanding of the issues and gain a larger representation of opinion.

- If a larger district or school team is assembled, coming to consensus about each item may be more difficult because people will bring not only different perspectives but also very different realities. For example, one person's school may deserve a high rating on one Action while another school in the district hasn't even considered that scope of work and therefore admittedly gets a lower mark. How can that be reconciled to reflect the work the district needs to accomplish? Or, the district may be doing well overall in one area but that work has not filtered into the schools. How should the team rate the district overall? There is likely to be rich, revealing discussion about many of the issues raised in the profile; staying focused on the status of the level of analysis (school or district) is essential.
- What one knows and doesn't know when asked to make judgments or evaluations influences one's answers to questions. In this activity—as in many others in this guide—participants' responses are directly related to their level of assessment literacy.

When you have rated all Actions and summarized the results, proceed with a team discussion of your current status, using the following questions as a springboard:

- Where are our strengths—places where our ratings seem high and we think that we have made real progress? What are the keys to our success on these fronts? List them.
- What have we accomplished to date? What is still to be done? Discuss specifics.
- Where are our omissions or weaknesses—areas where we have made little or no progress to date?
- How would we rank the Actions in terms of our progress? Rank them from 1 to 7, with 7 being most completely implemented.
- For areas of little progress to date, what have been our barriers? List them.
- How can we remove these barriers? Note suggestions.
- What should be our next priorities? Which pending Actions are most critical to our specific situation? How soon can/should we act on them?

Closure:

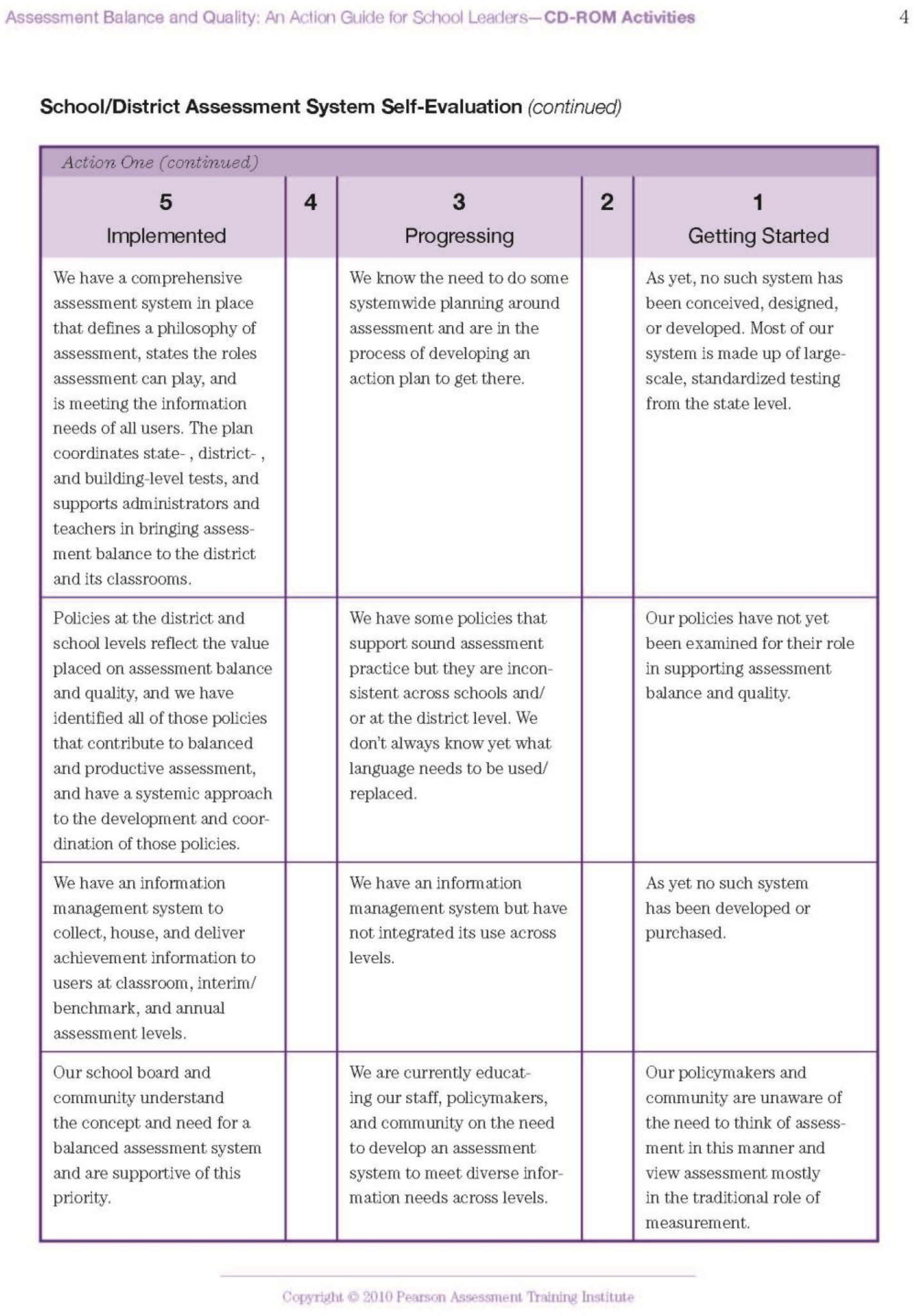
As we noted at the start of this Action Guide, our intention is to help you in two areas: (1) at a system/organizational (school or district) level; and (2) at a personal/professional level, one that considers the necessary knowledge and skills for leading assessment reform. We think it is helpful for teams to revisit this self-evaluation profile both before and after reading and doing many of the activities in Part 4. Doing the self-evaluation now will help clarify and increase understanding of the ten competencies for leaders you will encounter there. These competencies will be beneficial in implementing the Seven Actions. Coming back after reading Part 4 and reviewing the profile in light of these ten competencies will produce a deeper, more complete self-evaluation.

School/District Assessment System Self-Evaluation

Action One: Balance the district's assessment system to meet all key user needs

Balanced assessment systems blend effective assessment use at the classroom level with interim/benchmark assessment and annual testing to serve both formative and summative purposes. This Action urges examination of current levels of balance and movement toward greater balance if needed.

5 Implemented	4	3 Progressing	2	1 Getting Started
All faculty and staff are aware of differences in assessment purpose across classroom, interim/benchmark, and annual levels, and know how to use each to support and/or verify student learning; that is, to balance formative with summative assessment. We also understand what uses can and cannot be made with each level of assessment.		There is inconsistency among staff regarding assessment purpose, and some confusion about what is formative and what is summative. We are aware of the need for balance and have begun to plan for a balanced system.		There is little understanding of differences in purpose and assessment users, or appropriate uses of results across classroom, interim/benchmark, and annual levels.
A top assessment priority is to help students develop the capacity to assess their own learning and to use assessment results to help promote further learning.		Some faculty and staff recognize that students are important users of assessment information who make data-based instructional decisions that impact their own success, and have made some progress in helping them do so.		Students have not been viewed as key assessment users and there is little awareness of the benefits of bringing them into the assessment process, or knowledge of how to do so.



School/District Assessment System Self-Evaluation *(continued)*

Action One (continued)

5 Implemented	**4**	**3** Progressing	**2**	**1** Getting Started
We have a comprehensive assessment system in place that defines a philosophy of assessment, states the roles assessment can play, and is meeting the information needs of all users. The plan coordinates state- , district- , and building-level tests, and supports administrators and teachers in bringing assessment balance to the district and its classrooms.		We know the need to do some systemwide planning around assessment and are in the process of developing an action plan to get there.		As yet, no such system has been conceived, designed, or developed. Most of our system is made up of large-scale, standardized testing from the state level.
Policies at the district and school levels reflect the value placed on assessment balance and quality, and we have identified all of those policies that contribute to balanced and productive assessment, and have a systemic approach to the development and coordination of those policies.		We have some policies that support sound assessment practice but they are inconsistent across schools and/or at the district level. We don't always know yet what language needs to be used/replaced.		Our policies have not yet been examined for their role in supporting assessment balance and quality.
We have an information management system to collect, house, and deliver achievement information to users at classroom, interim/benchmark, and annual assessment levels.		We have an information management system but have not integrated its use across levels.		As yet no such system has been developed or purchased.
Our school board and community understand the concept and need for a balanced assessment system and are supportive of this priority.		We are currently educating our staff, policymakers, and community on the need to develop an assessment system to meet diverse information needs across levels.		Our policymakers and community are unaware of the need to think of assessment in this manner and view assessment mostly in the traditional role of measurement.

School/District Assessment System Self-Evaluation *(continued)*

Action One (continued)

5 Implemented	4	3 Progressing	2	1 Getting Started
We have inventoried all assessments used in the district and have categorized them by purpose, standards/targets measured, time of year, etc. for the purpose of understanding the balance we have in our current assessment system.		We are in the process of identifying all of the various assessments used at the district and school level for the purpose of getting a clearer understanding of what is currently in our assessment system.		We do not have a comprehensive picture of what assessments are currently being given.

Action Two: Refine achievement standards to reflect clear and appropriate expectations at all levels

Achievement standards are fundamental to any assessment system. That is, clear learning targets are needed to underpin classroom, interim/benchmark, and annual assessments. This Action calls for developing local achievement expectations as a foundation for balanced assessment.

5 Implemented	4	3 Progressing	2	1 Getting Started
We continue to refine our local achievement standards, have aligned them with state standards, and have identified our highest-priority learning outcomes.		We are aware of the need to develop clear local academic standards aligned to state standards and are in the process of doing so. What is in place is not yet used consistently across classrooms.		Local learning expectations are not in place.
Assessment results for all uses are always linked back to the local content standards.		We can link some assessments back to our written curriculum, but don't always know how or why we should do that.		We use the results as they are delivered to us and have yet to take the extra step of consistently matching results to the written curriculum.
We have deconstructed our standards into knowledge, reasoning, performance skills, and product development learning targets at each grade level for each subject.		We are in the process of deconstructing each of our standards into the scaffolding of grade-level curricula.		The deconstruction process has not been initiated.

School/District Assessment System Self-Evaluation *(continued)*

Action Two (continued)				
5 Implemented	**4**	**3** Progressing	**2**	**1** Getting Started
We have transformed the grade- and course-level learning targets that guide classroom assessment and instruction into student- and family-friendly versions.		Some of that work has been accomplished but we have not completed it for all grade levels and courses or it is not adequately communicated to parents and/or students.		We have yet to begin this process.
We have verified that each teacher in each classroom is master of the content standards that their students are expected to master. We provide professional support in content areas to teachers when needed.		We have identified contexts in which professional development is needed to ensure teacher competence in terms of our standards and that learning is underway.		There has been no investigation of teacher preparedness in their own content area(s).
All teachers in the district have received adequate training and ongoing support in developing their understanding of the written curricular documents. Teachers are given time to collaboratively plan lessons aimed at accomplishing grade-level/subject expectations.		We share curricular documents with our teachers. If there are questions about the new curriculum we address them, and provide some training at the beginning of the year in the understanding and use of those documents.		The curricular documents are available on request or are given to teachers when the documents have undergone revisions.
A curriculum implementation plan is in place to ensure consistency in achievement expectations across classrooms. Teachers are held accountable for teaching the written curriculum.		We recognize a need for a curriculum implementation plan to ensure the written curriculum is the taught curriculum, and have taken some steps to ensure that.		We have not ensured that there is consistency in achievement expectations across teachers. What is taught in each classroom in the same subject/grade level can differ widely.
Model/sample lessons and assessments, linked to the content standards, are available and used for professional development.		This is true for some subjects and grade levels.		We do not have this in our school/district.

School/District Assessment System Self-Evaluation *(continued)*

Action Three: Ensure assessment quality in all contexts to support good decision making

Because a variety of decisions are made based on assessment results, all assessments at classroom, interim/benchmark, and annual levels of use must yield dependable information about student achievement. This Action urges the evaluation of current assessments to verify quality.

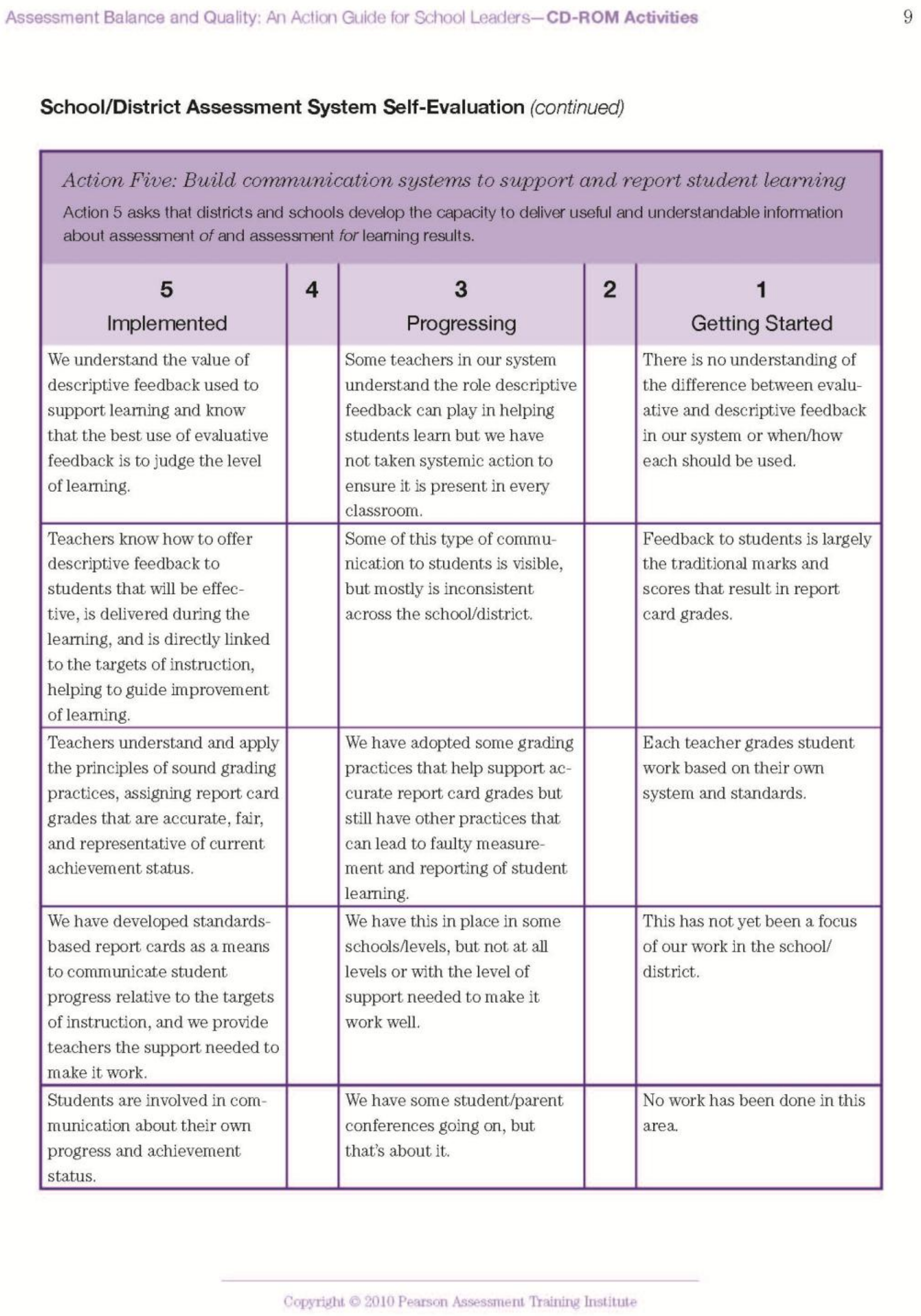
5 Implemented	4	3 Progressing	2	1 Getting Started
We have adopted and can apply the criteria by which we should judge the quality of our assessments, both *of* and *for* learning.		We have standards for assessment quality, and some district staff have the capability to evaluate for quality, but it is not a consistent condition in the district.		No such criteria have been identified; no quality control framework exists for us at any level.
There is general understanding that quality assessments form the foundation for accurate report card grades and for decisions made about students that rely on assessment data.		We subscribe to the use of multiple measures but haven't ensured that all data sources yield dependable results.		We've not considered this as a priority for our time/resources.
At the classroom level, teachers understand the importance of selecting the appropriate assessment method match to the type(s) of learning target to be assessed in order to help ensure quality results.		Teachers understand the need to vary assessment methods but may not apply strict quality criteria when doing so.		Teachers do not see the link between assessment quality and the assessment method used.
We have conducted a local evaluation of the quality of all of our assessments, including interim/benchmark and common assessments, if used.		We are aware of the need to conduct such an evaluation and are planning to conduct it.		There is no awareness of the need for or plans to conduct such an evaluation.

School/District Assessment System Self-Evaluation *(continued)*

Action Four: Help learners become assessors by using assessment for *learning strategies in the classroom*

By involving students in their own assessment during learning, teachers can maximize their confidence, motivation, and achievement. This Action urges that teachers involve them in assessment, understanding them as users of results just as they do themselves and others.

5 Implemented	4	3 Progressing	2	1 Getting Started
Faculty, staff, policymakers, and community members all understand and embrace the idea of assessment *for* learning—i.e., student-involved assessment to promote learning.		We are in the process of building local awareness of and belief in this set of ideas. Formative assessment is visible, but not as assessment *for* learning.		As yet, there is no awareness of the value of this concept or set of classroom practices.
Teachers use assessment information to focus instruction day to day in the classroom and communicate learning expectations to students in language they can understand.		Our primary use of formative assessment is at the interim or common assessment level, not exactly day-to-day at the classroom level. Some teachers know how to translate learning targets into student-friendly language, but many do not.		This has not been a focus or priority for us to date.
Teachers design assessments to help students self-assess and to help them use assessment results as feedback to set goals.		Some teachers administer assessments as practice; others need training to help them make that transition.		We don't involve students in the assessment process in these ways.



School/District Assessment System Self-Evaluation *(continued)*

Action Five: Build communication systems to support and report student learning

Action 5 asks that districts and schools develop the capacity to deliver useful and understandable information about assessment *of* and assessment *for* learning results.

5 Implemented	4	3 Progressing	2	1 Getting Started
We understand the value of descriptive feedback used to support learning and know that the best use of evaluative feedback is to judge the level of learning.		Some teachers in our system understand the role descriptive feedback can play in helping students learn but we have not taken systemic action to ensure it is present in every classroom.		There is no understanding of the difference between evaluative and descriptive feedback in our system or when/how each should be used.
Teachers know how to offer descriptive feedback to students that will be effective, is delivered during the learning, and is directly linked to the targets of instruction, helping to guide improvement of learning.		Some of this type of communication to students is visible, but mostly is inconsistent across the school/district.		Feedback to students is largely the traditional marks and scores that result in report card grades.
Teachers understand and apply the principles of sound grading practices, assigning report card grades that are accurate, fair, and representative of current achievement status.		We have adopted some grading practices that help support accurate report card grades but still have other practices that can lead to faulty measurement and reporting of student learning.		Each teacher grades student work based on their own system and standards.
We have developed standards-based report cards as a means to communicate student progress relative to the targets of instruction, and we provide teachers the support needed to make it work.		We have this in place in some schools/levels, but not at all levels or with the level of support needed to make it work well.		This has not yet been a focus of our work in the school/district.
Students are involved in communication about their own progress and achievement status.		We have some student/parent conferences going on, but that's about it.		No work has been done in this area.

School/District Assessment System Self-Evaluation *(continued)*

Action Six: Motivate students with learning success

The practice of relying on the anxiety and intimidation of accountability to motivate learning works for some students. It can energize those who have hope of success. But for students who have experienced chronic failure, turning up the anxiety will drive them more deeply into academic failure. For all students, a motivator that can work is success at learning. This Action urges educators to understand these emotional dynamics as they link assessment to student motivation and success.

5 Implemented	4	3 Progressing	2	1 Getting Started
Our faculty, staff, leaders, policymakers, and community understand the power student-involved assessment has to help all students experience the kind of academic success needed to remain motivated, confident, and engaged.		We are in the process of helping all stakeholders understand the motivational power of student-involved assessment *for* learning.		We largely motivate students by holding them accountable for learning.
The classroom assessment practices we use rely on student involvement in assessment during their learning to maintain their confidence and motivation.		The proportion of our teachers who involve their students in ongoing self-assessment as a motivator is increasing steadily.		Our classroom practices rarely include student-involved assessment as a motivator.

School/District Assessment System Self-Evaluation *(continued)*

Action Seven: Provide the professional development needed to ensure assessment literacy throughout the system

To successfully complete Actions 1–6 school districts must provide faculty and staff a foundational understanding of the principles of sound classroom assessment practice. This Action urges the provision of professional development in assessment literacy.

5 Implemented	4	3 Progressing	2	1 Getting Started
Leaders are committed to assessment literacy for all. Professional development resources have been allocated to achieve balance in our assessment systems, to have accurate assessments, and to employ assessment *for* learning practices.		We have begun to make school improvement and resource allocation decisions that reflect a desire to offer the professional development needed to form the foundation of a quality, balanced assessment system.		Such professional development is not yet a priority on our district.
Our school leaders have developed the assessment literacy they need to maintain the vision, to develop essential infrastructure, and support teacher development in assessment literacy.		We acknowledge the need to have all leaders assessment literate and leaders are finding opportunities to increase their knowledge and skills in quality, balanced assessment practices.		Assessment literacy has not been a focus of our development of school leaders.
The development of assessment literacy is offered in a professional development model that allows teachers to learn from each other in collaborative teams and practice in the classroom as they learn.		We have some teacher-directed, job-embedded staff development, but our system does not have the structures in place to support this kind of adult learning.		Our professional development model is still largely workshop based.
Professional development is having its desired impact as our program evaluation shows that we have achieved balance, a high degree of quality assessment, and an increase in student achievement.		Professional development appears to be working but we have little hard data to support that conclusion.		We are not evaluating our programs in ways that would tell us that what we do delivers results.

APPENDIX E

SURVEY OF MATH ASSESSMENT PRACTICES

RATING SCALE: 1=not helpful 2=somewhat helpful 3=very helpful

Formative assessments *provide continuous feedback during the teaching-learning cycle with the goal of modifying instruction. Examples include NWEA, non-graded quizzes, and fluency measures. Which formative assessment do you use (Please provide a brief description if this is an assessment that you have created)? To what extent are these assessments helpful to you in modifying your instruction?*

Summative assessments *document learning at the end of the teaching-learning cycle with the goal of documenting a level of mastery. Examples include MCAS, quizzes and tests. Which summative assessments do you use? To what extent are these assessments helpful to you in documenting your students' level of mastery?*

RATING SCALE: 1=not helpful 2=somewhat helpful 3=very helpful

Evaluative assessment *entails the systematic use of assessment to gauge the value, effectiveness or efficiency of an educational program. Examples include MCAS and other assessments that a teacher or district may design to measure the effectiveness of their curriculum or instructional practices. Which evaluative assessments do you use? To what extent are these assessments helpful to you in evaluating the effectiveness of your curriculum or instructional practices?*

APPENDIX F

INTERVIEW QUESTIONS FOR TEACHERS FOR QUESTION #2

For 6th Grade Teachers

1. What role do you have in this school system?

2. How long have you worked in that capacity?

3. What does a balanced system of math assessment mean to you?

4. Which types of math assessments do you use in your classroom?

5. How do the results from other math assessments get communicated to you?

6. What opportunities do you have to meet with teachers at your grade level to discuss math assessments or results?

7. What opportunities do you have to meet with teachers at the next grade level to discuss math assessments or results?

8. What assessment results from your current range of assessments are most helpful to you as you teach your current students?

9. What assessment results from your current range of assessments are most helpful to you as you try to inform next year's receiving teacher about your current students?

10. If you could change one thing about the current system of math assessments, what would that be?

11. Is there anything that we should have talked about but didn't?

For 7th Grade Teacher

1. What role do you have in this school system?

2. How long have you worked in that capacity?

3. What does a balanced system of math assessment mean to you?

4. Which types of math assessments do you use in your classroom?

5. How do math assessment results get communicated to you?

6. What opportunities do you have to meet with teachers at the 6th grade level to discuss math assessments or results?

7. What assessment results from your current range of assessments are most helpful to you as you try to place incoming 6th grade students into 7th grade classes:

- From within this district?
- From other Massachusetts districts?
- From out-of-state?

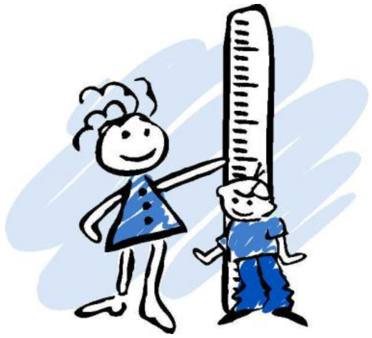
8. If you could change one thing about the current system of math assessments, what would that be?

9. Is there anything that we should have talked about but didn't?

APPENDIX G

ASSESSMENT LITERACY INVENTORY

Assessment Literacy Inventory



Cynthia Campbell, Ph.D.
Northern Illinois University

and

Craig A. Mertler, Ph.D.
Bowling Green State University

<u>Description of the *ALI*</u>:

The ***Assessment Literacy Inventory*** (***ALI***) consists of five scenarios, each followed by seven questions. The items are related to the seven "Standards for Teacher Competence in the Educational Assessment of Students." Some of the items are intended to measure general concepts related to testing and assessment, including the use of assessment activities for assigning student grades and communicating the results of assessments to students and parents; other items are related to knowledge of standardized testing, and the remaining items are related to classroom assessment.

<u>Directions</u>:

Read each scenario followed by each item carefully; select the response you think is the best one and *mark your response on the answer sheet*. Even if you are not sure of your choice, but, *mark the response you believe to be the best.*

Scenario #1

Ms. O'Connor, a math teacher, questions how well her 10th grade students are able to apply what they have learned in class to situations encountered in their everyday lives. Although the teacher's manual contains numerous items to test understanding of mathematical concepts, she is not convinced that giving a paper-and-pencil test is the best method for determining what she wants to know.

1. Based on the above scenario, the type of assessment that would *best* answer Ms. O'Connor's question is called a/an
   A. performance assessment.
   B. authentic assessment.
   C. extended response assessment.
   D. standardized test.

2. In order to grade her students' knowledge accurately and consistently, Ms. O'Connor would be well advised to
   A. identify criteria from the unit objectives and create a scoring rubric.
   B. develop a scoring rubric after getting a feel for what students can do.
   C. consider student performance on similar types of assignments.
   D. consult with experienced colleagues about criteria that has been used in the past.

3. To get a general impression of how well her students perform in mathematics in comparison to other 10th graders, Ms. O'Connor administers a standardized math test. This practice is acceptable *only* if
   A. the reliability of the standardized test does not exceed .60.
   B. the standardized test is administered individually to students.
   C. the content of the standardized test is well known to students.
   D. the comparison group is comprised of grade level peers.

4. Which of the following is an *in*appropriate use of the results from this standardized math test?
   A. planning instruction
   B. assigning student grades
   C. determining students' strengths and weaknesses
   D. developing curriculum

5. Throughout instruction, Ms. O'Connor assesses how well her students are grasping the material. These assessments range from giving short quizzes following introduction to a new topic, to administering an end-of-the-unit final exam. In order to improve the validity of this grading procedure, Ms. O'Connor should
   A. make the grading scale the same for all assessments.
   B. consider students' prior performance before assigning a final grade.
   C. weight assessments according to their relative importance.
   D. take into consideration each student's effort when calculating grades.

6. During a parent teacher conference, one of the parents of a student in Ms. O'Connor's class wants to know what it means that his daughter scored in the 80th percentile in mathematics. Which of the following provides the *best* explanation of this student's score?
   A. She got 80% of the items on the math test correct.
   B. She is likely to earn a grade of 'B' in her math class.
   C. She is demonstrating above grade level performance in math.
   D. She scored the same or better than 80% of the norm group.

7. Which of the following is an appropriate use of assessment information?
   A. Utilize information from a variety of assessments when making decisions about student learning.

B. Use scores from standardized tests to determine teacher instructional effectiveness.
C. Use scores from a standardized test as the primary indicator of student retention.
D. Post final grades in order to provide normative information to students in the class.

Scenario #2

Mr. Okawa, a 5th-grade teacher, is planning his instruction for the next grading period, aware of the fact that his students will be taking the statewide achievement test near the end of the grading period.

8. Mr. Okawa's mathematics unit for this grading period will focus on multi-step problem-solving. He wants to assess his students' problem-solving abilities at the end of the unit to determine if any reinstruction will be necessary prior to the statewide test. Which of the following assessment strategies would be the most appropriate choice?
   A. He should choose the assessment included in the teacher's manual from the textbook he uses.
   B. He should choose an assessment which is consistent with the content and skills he taught.
   C. He should choose a different standardized assessment that provides a score on similar skills.
   D. He should choose an assessment which covers single-step problem-solving skills.

9. Mr. Okawa decides to develop his own assessment in order to determine if any reinstruction will be necessary. He also wants to use his assessment as a means of anticipating how his students will perform on the statewide assessment. In order for him to accurately approximate his students' performance, which of the following would be the most appropriate type of assessment for him to develop?
   A. a performance assessment
   B. a multiple-choice test
   C. a portfolio assessment
   D. an essay test

10. Julie, one of Mr. Okawa's students, receives a percentile rank of 60 on the problem-solving skills subtest of the statewide assessment. This score is most appropriately interpreted as which of the following?
   A. Julie scored above average.
   B. Julie scored below average.
   C. Julie scored at the national average.
   D. Not enough information to determine.

11. Juan, another student in Mr. Okawa's class, receives a scaled score of 196 on the reading comprehension portion of the statewide assessment. The cut score is 200; therefore, Juan does not pass this subtest. However, the subtest has a standard error of measurement equal to 6. Which of the following is the best decision for Mr. Okawa to make regarding instruction appropriate to meet Juan's needs?
   A. Juan has clearly not achieved the minimum level of reading comprehension and should receive remedial reading instruction.
   B. Mr. Okawa knows that Juan could have scored higher, so the results of the test should be ignored.
   C. Juan may likely have achieved the minimum level of reading comprehension and nothing different or additional should be done.
   D. Mr. Okawa knows that Juan should have scored much lower, so the results of the test should be ignored.

12. Which grading practice being considered by Mr. Okawa would result in grades that would least reflect achievement?
   A. grades based on daily homework and chapter tests
   B. grades based on daily homework and chapter tests, with points deducted for poor effort
   C. grades based on daily homework and chapter tests, where students are permitted

to redo assignments in order to meet higher standards
D. grades based on chapter tests, where daily homework is not formally graded

13. Barbara scores at the 60th percentile on mathematics problem-solving and at the 56th percentile on reading comprehension. The percentile bands for each test are five percentile ranks wide. What advice should Mr. Okawa give to Barbara's parents?
    A. They should ignore the difference; her performance was essentially the same on the two tests.
    B. They should seek additional tutoring help for Barbara in reading.
    C. They should force Barbara to read more at home.
    D. They should provide enrichment experiences for Barbara in math, which is her better performance area.

14. Mr. Okawa was worried that his students would not perform well on the statewide assessment. He did all of the following to help increase students' scores. Which was unethical?
    A. He instructed students in strategies for taking multiple-choice tests, such as how to use answer sheets.
    B. He planned his instruction so that it focused on concepts and skills to be covered on the test.
    C. He encouraged the students to do their best, and provided them with a reward after testing was complete.
    D. He allowed students to practice with items from an alternate form of the test.

<u>Scenario #3</u>

Ms. Green is an 8th-grade American History teacher. She has just finished teaching a unit on the Industrial Revolution and wishes to make decisions about her students regarding their higher-order thinking skills. Ms. Green has decided to give her students a single assessment in the form of an end-of-unit multiple-choice test. She anticipates that most of her students will perform well on the test.

15. Based on her goal, what can you conclude about her decision to administer a multiple-choice test?
    A. This is an appropriate choice for a unit assessment.
    B. The test scores may not be valid for this purpose.
    C. The test scores may not be reliable for this purpose.
    D. A true-false test would be more appropriate.

16. To determine the quality of her multiple-choice test, Ms. Green should conduct an item analysis and examine all of the following except
    A. item difficulty values.
    B. item discrimination values.
    C. reliability coefficients.
    D. validity coefficients.

17. Ms. Green decides to score the tests using a 100-percent correct scale. Generally speaking, what is the proper interpretation of a student score of 85 on this scale?
    A. The student answered 85% of the items on the test correctly.
    B. The student knows 85% of the content covered by this instructional unit.
    C. The student scored higher than 85% of other students who took this test.
    D. The student scored lower than 85% of other students who took this test.

18. Some of Ms. Green's students do not score well on the multiple-choice test. She decides that the next time she teaches this unit, she will begin by administering a pretest to check for students' prerequisite knowledge. She will then adjust her instruction based on the pretest results. What type of information is Ms. Green using?
    A. norm-referenced information
    B. criterion-referenced information
    C. both norm- and criterion-referenced information
    D. neither norm- nor criterion-referenced information

19. The Industrial Revolution test is the only student work that Ms. Green grades for the current grading period. Therefore, grades are assigned only on the basis of the test. What is the major criticism of this practice?
    A. The test, and therefore the grades, reflect too narrow a curricular focus.
    B. These grades, since based on tests alone, is probably biased against some minority students.
    C. She should add extra points to the scores of students who scored low on the test.
    D. Decisions like grades should be based on more than one piece of information.

20. Mr. Simpson, another American History teacher, bases his grades primarily on his observations of students during class. The primary distinction between his system of assigning grades and that used by Ms. Green is *best* characterized as which of the following?
    A. Ms. Green uses formal assessment; Mr. Simpson uses informal assessment.
    B. Ms. Green uses formative assessment; Mr. Simpson uses summative assessment.
    C. Ms. Green uses standardized assessment; Mr. Simpson uses nonstandardized assessment.
    D. Ms. Green uses traditional assessment; Mr. Simpson uses alternative assessment.

21. Based on their grades from last year, Ms. Green believes that some of her low-scoring students are brighter than their test scores indicate. Based on this knowledge, she decides to add some points to their test scores, thus raising their grades. Which of Ms. Green's actions was unethical?
    A. examining her student's previous academic performance
    B. adjusting grades in her course
    C. using previous grades to adjust current grades
    D. adjusting some students' grades and not others'

<u>Scenario #4</u>

Mr. Valdez is an English teacher in the newly built middle school. Experienced in issues of classroom assessment, Mr. Valdez is often asked to respond to the district's questions concerning best practices for evaluating student learning.

22. Ms. Franklin, also an English teacher, asks what type of assessment is best for evaluating her 6th graders' writing skills. Which of the following methods is likely to provide the *best* response to her question?
    A. selected response methods
    B. true/false statements
    C. completion items
    D. essay prompts

23. One of the middle school math teachers is redesigning her tests to make greater use of "story problems" as a way to check students' math understanding. She consults with Mr. Valdez to see what, if any, concerns she should be aware of when constructing assessments of this type. Which statement is *not* an appropriate recommendation when designing story-based math tests?
    A. make sure that the reading level is grade appropriate
    B. avoid scenarios more familiar to certain groups over others
    C. check for clarity of sentence construction
    D. incorporate scenarios used during instruction

24. Isabel, a student in Mr. Valdez's class, scored 78 points on a standardized English test which had a mean of 80 and a standard deviation of 4. She scored 60 points on the science portion of this test which had a mean of 50 and a standard deviation of 3. Based on the above information, in comparison to her peers, which statement provides the most accurate interpretation?
    A. Isabel is better in English than in science.
    B. Isabel is better in science than in English.
    C. Isabel is below average in both subjects.

D. Isabel is close to average in both subjects.

25. At the end of each class period, Mr. Valdez does a quick “check in” with his students to get an impression of their understanding. In this example, the primary purpose for conducting formative assessment is to
A. identify cumulative knowledge.
B. determine content for the final exam.
C. plan classroom instruction.
D. evaluate curriculum appropriateness.

26. To prepare students for state testing and identify areas of school improvement, all 6th grade English teachers give a common final exam which contains a series of essay items. Recently, however, several teachers have expressed concern that the time and effort necessary to complete grading on a timely basis may result in inconsistent scoring. They consult with Mr. Valdez. Which of the following provides the *best* response to the teachers’ concern for consistency?
A. grade all responses to essay #1 before grading responses to essay #2
B. during grading, adjust rubric criteria to reflect exemplary student work
C. utilize a holistic scoring method to minimize teacher subjectivity in scoring
D. all things being equal, it is best to limit the use of multiple essay exams

27. Jeremy, a 6th grade student in Mr. Valdez’s class, received a grade equivalent score of 7.2 on a standardized reading test. Jeremy’s parents wonder what this means. Based on the above information, which of the following statements provides the most appropriate interpretation of this student’s score?
A. Jeremy is reading at the 7th grade level.
B. Jeremy is reading better than the majority of students in his class.
C. Jeremy is reading 6th grade material as expected.
D. Jeremy should be placed in a 7th grade reading class.

28. “To ensure that standardized test results provide an accurate picture of what students *re*
know, it is recommended that teachers clarify items that are confusing to students.”

Based on best practices of assessment, which of the following is an appropriate response to the above statement?
A. The above statement is an acceptable way to reduce error in testing.
B. The above statement is an acceptable way to increase test validity.
C. The above statement is unacceptable because it labels students as poor readers.
D. The above statement is unacceptable because it breaks standardization.

<u>Scenario #5</u>

Ms. Hawkins is responsible for teaching science at the 4th grade level. Over the past couple of years, her students have really seemed to struggle with investigations of how water changes from one state to another (i.e., freezing, melting, condensing, and evaporating), but she is unsure of where the specific difficulties lie. She is aware that her students need to improve their conceptual understanding of this content standard.

29. Ms. Hawkins wishes to conduct some sort of assessment in order to identify the specific difficulties her students are experiencing. Which of the following would *best* meet her needs?
A. a diagnostic assessment
B. an informal assessment
C. a standardized assessment
D. a summative assessment

30. In an effort to refine both her instruction and assessment of this content, Ms. Hawkins conducts an item analysis of student scores from last year’s final unit test over this material. She should definitely discard or substantially revise a test item that
A. has a difficulty value between .50 and .75.
B. has a discrimination value equal to +.30.
C. has a discrimination value equal to -.50.
D. has a difficulty value equal to .90.

31. Ms. Hawkins' unit test also includes a restricted-response essay item. She is concerned with the demonstrated level of understanding of several specific criteria in her students' responses. Which of the following would *best* facilitate her scoring of these responses?
    A. an objective answer key
    B. a holistic rubric
    C. a checklist
    D. an analytic rubric

32. Following the completion of the unit, Ms. Hawkins determines that her students have satisfactorily mastered these concepts. However, when her students take the statewide standardized assessment in the spring, she notices that her students perform very poorly on items addressing these same concepts. Considering the discrepancy between students' classroom performance and their standardized test results, what action is most appropriate when making decisions concerning school improvement?
    A. recommend that classroom instruction be consistent among 4th grade science teachers
    B. ensure alignment between instruction and what is measured on the standardized test
    C. select a standardized test that is more likely to yield higher scores in science
    D. identify the percentage of students predicted to perform well in advanced science classes

33. Ms. Hawkins wants to be sure that the term grades she assigns to her students' performance in science reflect each student's respective level of content mastery for that unit. Which of the following grading systems would *best* accomplish this goal?
    A. a criterion-referenced grading system
    B. a norm-referenced grading system
    C. a pass–fail grading system
    D. a portfolio grading system

34. Nolan is a student in Ms. Hawkins' class. He receives a raw score of 12 items answered correctly out of a possible 15 on the physical science portion of a standardized test. This raw score equates to a percentile rank of 45. His parents are confused about how he could answer so many items correctly, but receive such a low percentile rank. They approach Ms. Hawkins for a possible explanation. Which of the following is the appropriate explanation to offer to his parents?
    A. "I don't know…there must be something wrong with the way the test company figured the scores."
    B. "Although Nolan answered 12 correctly, numerous students answered more than 12 correctly."
    C. "Raw scores are purely criterion-referenced and percentile ranks are merely one form of norm-referenced scoring."
    D. "Raw scores are purely norm-referenced and percentile ranks are merely one form of criterion-referenced scoring."

35. In an attempt to try to encourage and motivate her students who are struggling academically, Ms. Hawkins decides to share her gradebook, especially test scores, with them in order to demonstrate how well others are performing. Another teacher advises her not to do this, as it is a clear violation of
    A. *The Code of Fair Testing Practices in Education.*
    B. *The Family and Education Rights and Privacy Act.*
    C. *The Standards for Teacher Competence in the Educational Assessment of Students.*
    D. *The No Child Left Behind Act.*

Assessment Literacy Inventory
Scoring Key

1.	B
2.	A
3.	D
4.	B
5.	C
6.	D
7.	A
8.	B
9.	B
10.	A
11.	C
12.	B
13.	A
14.	D
15.	B
16.	D
17.	A
18.	B
19.	D
20.	A
21.	D
22.	D
23.	D
24.	B
25.	C
26.	A
27.	C
28.	D
29.	A
30.	C
31.	D
32.	B
33.	A
34.	B
35.	B

Alignment of Standards with items on *ALI*:

Standard 1—Items 1, 8, 15, 22, 29
Standard 2—Items 2, 9, 16, 23, 30
Standard 3—Items 3, 10, 17, 24, 31
Standard 4—Items 4, 11, 18, 25, 32
Standard 5—Items 5, 12, 19, 26, 33
Standard 6—Items 6, 13, 20, 27, 34
Standard 7—Items 7, 14, 21, 28, 35

REFERENCES

Abrams, L. A. (2007). Implications for high-stakes testing for the use of formative classroom assessment. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 79-98). New York, NY: Teachers College Press.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). Standards for teacher competence in the educational assessment of students. *Educational measurement: Issues and practice, 9* 4), 30-32.

Andrade, H., & Boulay, B. (2003). Role of rubric-referenced self-assessment in learning to write. *The Journal of Educational Research, 97*(1), 21-34.

Association of Assessment in Counseling. (2003). Standards of Multicultural Assessment. Retrieved from http://aac.ncat.edu/Resouurces/documents/StandardsforMulticulturalAssesment.pdf

Black, P. (1993a). Assessment policy and public confidence: Comments on the BERA Policy Task Group's article "Assessment and the improvement of education." *The Curriculum Journal, 4*(3), 421-427.

Black, P. (1993b). Formative and summative assessment by teachers. *Studies in Science Education, 21,* 49-97.

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

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment Evaluation Accountability, 21*, 5-31.

Bloom, B., Hastings, J. T., & Madaus, G. (Eds.). (1971). *Handbook on formative and summative evaluation of student learning*. New York, NY: McGraw-Hill Book.

Boudett, K., City, E., & Murnane, R. (2005). *Data wise*. Cambridge, MA: Harvard Education Press.

Boudett, K., & Steele, J. (2007). *Data wise in action.* Cambridge, MA: Harvard Education Press.

Brookhart, S. (1999). Teaching about communicating assessment results and grading. *Educational Measures: Issues and Practices, 18,* 5-13.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist, 24*, 409-429.

Campbell, C., Murphy, J., & Holt, J. (2002, October). *Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers.* Paper presented at eh annual meeting of the Mid-Western Educational Research Association. Columbus, OH.

Capizzi. A., & Fuchs, L. (2005). Effects of curriculum-based measurement with and without diagnostic feedback on teacher planning. *Remedial and Special Education, 26*(3), 159-174.

Chappuis, J. (2009). *Seven strategies of assessment for learning*. Boston, MA: Pearson Education.

Chappuis, S., Commodore, C., & Stiggins, R. (2010). *Assessment balance and quality: An action guide for school leaders.* Portland, OR: Assessment Training Institute.

Chen, P. (2005). Teacher candidates' literacy in assessment. *Academic Exchange*, (Fall), 62-66.

Cizek, G. J. (2007). Formative assessment and large-scale assessment: Implications for future research and development. In J. H. McMillan (Ed.), *Formative classroom assessment* (pp. 99-115). New York, NY: Teachers College Press.

Clarke, B., & Shinn, M. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review, 33*(4), 234-248.

Council of Chief State School Officers (2002). *Making valid and reliable decisions in determining adequate yearly progress*. Retrieved http://programs.ccsso.org/content/pdfs/AYPpaper.pdf

Council of Chief State School Officers (2004). *A framework for examining validity in state accountability systems.* Retrieved http://www.ccsso.org/Documents/2004/Framework_For_Examining_Validity_2004.pdf

Council of Chief State School Officers (2008). *Educational leadership policy standards: ISLLC 2008*. Retrieved http://www.wallacefouundation.org/Knowledge Center/Knowledge Topics/Current Areas of Focus/Educational Leadership

Creswell, J. (2003) *Research design*, Thousand Oaks, CA: Sage.

Crooks, T. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438-481.

Cross, C. T. (2004). Two Bushes and a Clinton. *Political education: National policy comes of age* (pp. 91-125). New York, NY: Teachers College Press.

Deno, S. (1985) Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219-232.

DuFour, R., Eaker, R., & Dufour, R. (2005). On common ground: The power of professional learning communities. Bloomington, IN: Solution Tree.

Durden, T. (2008). Do your homework! Investigating the role of culturally relevant pedagogy in comprehensive school reform models serving diverse student populations. *The Urban Review, 40,* 403-419.

Dweck, C. (1986). Motivational processes affecting learning. *American Psychologist, 41*(special issue), 1040-1048.

Earl, L. (2003). *Assessment as learning*. Thousand Oaks, CA: Corwin Press.

Earl, L., & Timperley, H. (2009). Understanding how evidence and learning conversations work. In L. M. Earl & H. Timperley (Eds.), *Professional learning conversations: Challenges in using evidence for improvement* (pp. 1-12). Springer Science and Business Media.

Elementary and Secondary Education Act of 1965. (1965). Pub. L. No. 89-100; 79 Stat. 27.

Fernandez, M., & Fontana, D. (1996). Changes in control beliefs in Portuguese primary school pupils as a consequence of the employment of self-assessment strategies. *British Journal of Educational Psychology, 66*, 301-313.

Fontana, D., & Fernandez, M. (1994). Improvements in Mathematics performance as a consequence of self-assessment in Portuguese primary school pupils. *British Journal of Educational Psychology, 64,* 407-417.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative: A meta-analysis. *Exceptional Children, 53*(3), 199-208.

Fuchs, L. S., & Fuchs, D. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22*(1), 22-49.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in Mathematics operations. *American Educational Research Journal, 28*(3), 617-641.

Gewertz, C., & Robelen, E. (2010). U.S. tests awaiting big shifts. *Education Week, 30*(3), pp. 1, 18 & 19.

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

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

Hamilton, L., Stecher, B., & Yuan, K. (2008). Standards-based reform in the united states: history, research and future directions (pp. 1-76). In Center on Education Policy (Ed.), *Rethinking the federal role in education*. Washington, DC: Author.

Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education, 4*(3), 365-379.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Hintze, J. M., Christ, T. & Methe, S. (2006). Curriculum-based measurement. *Psychology in Schools, 43*(1), 45-56.

House, E. R. (1993). *Professional Evaluation*. Newbury Park, CA: Sage.

Huai, N., Braden, J., White, J., & Elliott, S. (2006). Effect of an internet-based professional development program on teachers' assessment literacy for all students. *Teacher Education and Special Education*, *29*(4), 244-260.

Improving America's Schools Act of 1994. (1994). Pub. L. No. 103-382; 108 Stat. 3518.

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

Kulik, C.-L., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research, 60*(2), 265-299.

Lasky, S., Schaffer, G., & Hopkins, T., (2009). Learning to think and talk from evidence: Developing system-wide capacity for learning conversations. In L. M. Earl & H. Timperley (Eds.), *Professional learning conversations: Challenges in using evidence for improvement* (pp. 95-107). Springer Science and Business Media.

Love, N., Stiles, K., Mundry, S., & DiRanna, K. (2008). *The data coach's guide to improving learning for all students.* Thousand Oaks, CA: Corwin Press.

Lukin, L., Bandalos, D., Eckhout, T., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice*, *23*, 26-32.

Manna, P. (2004). Leaving no child behind. In C. T. Cross (Ed.), *Political education: National policy comes of age*. (pp. 126-143). New York, NY: Teachers College Press.

Massachusetts Department of Elementary and Secondary Education (2010). *Race to the top*. Malden, MA: Author.

Massachusetts Department of Elementary and Secondary Education. (2011). *Massachusetts curriculum frameworks*. Retrieved from http://www.doe.mass.edu/frameworks/current.html

Massachusetts Department of Elementary and Secondary Education. (2012). *Massachusetts granted flexibility from portions of No Child Left Behind Act to focus on innovative methods for ensuring all students achieve at high levels*. Retrieved from http://www.doe.madd.edu/news/news.aspx?id=6666

McDonald, B., & Boud, D. (2003). The impact of self-assessment on achievement: The effects of self-assessment training on performance in external examinations. *Assessment in Education*, (10), 209-220.

McMillan, J. (2007). Formative classroom assessment: The key to improving student achievement. In J. H. McMillan (Ed.), *Formative classroom assessment* (pp. 1-7). New York, NY: Teachers College Press.

Mertler, C. (2003). *Preservice versus inservice teachers' assessment literacy: does classroom experience make a difference?* Paper presented at the annual meeting of the Mid-Western Educational Research Association. Columbus, OH.

Mertler, C., & Campbell, C. (2005). *Measuring teachers' knowledge and application of classroom assessment concepts: Development of the assessment literacy inventory*. Paper presented at the annual meeting of the American Educational Research Association. Montreal, Quebec, Canada.

McDermott, K. (2011a). *High-stakes reform*. Washington, DC: Georgetown University Press.

McDermott, K. (2011b). *Interstate governance of standards and testing*. Paper presented at the Rethinking education governance in the 21st century. Thomas B. Fordham Institute and the Center for American Progress, Washington, DC.

Millitello, M., Schweid, J., & Sireci, S. (2010). Formative assessment systems: Evaluating the fit between school districts' needs and assessment systems' characteristics. *Educational Assessment Evaluation Accountability,* (22), 29-52.

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist, 22*(2), 155-175.

National Commission on Excellence in Education (1983) *A Nation at Risk: The Imperative for Educational Reform.* Washington, D.C: United States Department of Education.

No Child Left Behind Act. (2002). Pub. L. No. 107-110; 114 Stat. 1425.

Olson, L. (1995). Cards on the table. *Education Week*, *14,* 23-28.

Partnership for Assessment of Readiness for College and Careers (2012*). Common core implementation workbook.* Retrieved http://www.parcconline.org/implementation

Patton, M. (2008). *Utilization-focused evaluation*. Thousand Oaks, CA: Sage.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know.* Washington, DC: National Academy Press.

Pellegrino, J. W., & Goldman, S. (2008). Beyond rhetoric: Realities and complexities of integrating assessment in classroom teaching and learning. In C. Dwyer (Ed.), *The future of assessment* (pp. 7-52). New York, NY: Taylor & Francis.

Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The Role of interim assessments in a comprehensive assessment system*. Aspen, CO: Center for Assessment.

Plake, B. S., & Impara, J. C. (1993). *Teacher assessment literacy questionnaire*. University of Nebraska-Lincoln.

Popham, W. J. (2006). *Assessment for educational leaders.* Boston, MA: Pearson Education.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental. *Theory into Practice, 48*, 4-11.

Rothman, R. (2010). *Principles for a comprehensive assessment system*. Washington, DC: Alliance for Excellent Education.

Russ-Eft, D., & Preskill, H. (2009). *Evaluation in organizations: A systemic approach to enhancing learning, performance and change.* New York, NY: Perseus Books Group.

Ryan, K. (2002). Shaping educational accountability systems. *American Journal of Evaluation*, *23*(4), 453-468.

Sadler, D. R. (1983). Evaluation and the improvement of academic learning. *Journal of Higher Education, 54*(1), 60-79.

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

Sargent, J. L. (2003). *The data retreat facilitator's notebook* (CESA #7). Green Bay, WI: Cooperative Educational Service Agency.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives on curriculum evaluation* (pp. 39-83). Chicago, IL: Rand McNally.

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

Shepard, L. (2008). Formative assessment: Caveat emptor. In C. Dwyer (Ed.), *The future of assessment* (pp. 279-303). New York, NY: Taylor & Francis.

Shepard, L., Hammerness, K., Darling-Hammond, L., Rust, F., Snowden, J., Gordon, E., Gutierrez, C., & Pacheco, A. (2005). In L Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world* (pp. 275-326). San Francisco, CA: Wiley.

Smith, C., & Freeman, R. (2002). Using continuous system level assessment to build school capacity. *American Journal of Evaluation, 23*(3), 307-319.

Stiggins, R. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77*(3), 238-245.

Stiggins, R. (2001). *Student-involved classroom assessment* (3rd ed.). Columbus, OH: Merrill Prentice-Hall.

Stiggins, R., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, *22*(4), 271-286.

Taras, M. (2005). Assessment-summative and formative-Some theoretical reflections. *British Journal of Educational Studies, 53*(4), 466-478.

Thurber, R., Shinn, M., Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review, 31*(4), 498-513.

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

Wang, L., Beckett, G., & Brown, L. (2006). Controversies of standardized assessment in school accountability reform: A critical synthesis of multidisciplinary research evidence. *Applied Measurement in Education, 19*(4), 305-328.

Wiggins, G., & McTighe, J. (2007). *Schooling by design*. Alexandria, VA: Association for Supervision and Curriculum Development.