Research Watch
WAKE COUNTY PUBLIC SCHOOL SYSTEM
EVALUATION AND RESEARCH DEPARTMENT

E&R Report No. 11.10 **June 2011**

COMPREHENSIVE ASSESSMENT SYSTEMS: PURPOSES AND IMPLEMENTATION

Talbot Troy

INTRODUCTION

In 2007, Rhode Island's Providence Public School District (PPSD) discontinued a benchmark (a.k.a. interim) assessment program that had been developed by district personnel in 2004. PPSD administrators reported that, although the program was largely successful, it was discontinued due to high costs. A closer examination reveals that the program may also have been plagued by other problems common in districts throughout the United States (Clune & White, 2008). For one thing, much of the professional development effort was dedicated to training teachers to access and read data reports, rather than analyzing student proficiency and improving instruction. Also, the PPSD benchmark assessments were cast in a dual role as benchmark *and* formative, an arrangement that threatens to erode both purposes by blurring important lines of distinction (Goertz, Olah, & Riggan, 2010; Herman, Osmundson & Dietel, 2010; Perie, Marion, & Gong, 2009).

Indeed, confusion between benchmark and formative assessment is widespread and problematic (Chappuis & Chappuis, 2008; Crane, 2008; Goren, 2010; Moss & Brookhart, 2009; Popham, 2008; Sharkey & Murnane, 2006). Part of this is due to marketing practices of vendors who use the term "formative" in reference to instruments previously packaged as test preparation materials (Li, Marion, Perie, & Gong, 2010; Olson, 2005; Popham, 2008; Shepard, 2010). Confusion may also arise from inconsistencies in current literature meant to provide clarity. For example, Sharkey and Murnane (2006) do not define either formative or benchmark assessment but use the phrase "formative assessment system" to describe district-designed assessments coupled with a data collection system to be used by central administrators and teachers alike—characteristics experts attribute to benchmark assessment.

While there is direct evidence of the effectiveness of formative assessment to improve student learning (Black & Wiliam, 1998), studies have been inconclusive regarding the effect of

The author would like to acknowledge the support and intellectual contributions from David Holdzkom, Bradley McMillen, and Sonya Stephens.

benchmark assessment (Goertz, et al., 2010; Henderson, Petrosino, Guckenburg, & Hamilton, 2008; Shepard, 2010). In fact, some writers caution that in mimicking state-wide standardized tests, benchmark assessments may compound current trends towards "knowledge-lean and process-constrained" instruction (Pellegrino, 2004, p. 9; Shepard, 2010). However the summative data sought by administrators can not be expected to emerge from the practice of formative assessment (Li, et al., 2010; Popham, 2008). In addition, there are indications that certain types of benchmark assessment systems support desirable classroom and collegial practices of teachers, when adequate instructional leadership is provided at the school level (Bulkley, Olah, & Blanc, 2010; Crane, 2008; Downey, Steffy, Poston, & English, 2009; Goertz, et al., 2010; Kerr, Marsh, Ikemoto, Darilek, & Barney, 2006). For example, PPSD's program seems to have resulted in better alignment of instruction with curriculum, and it increased the use of data by teachers for collaboration about instructional decisions.

For any district developing a comprehensive assessment plan, there may be valuable lessons in the report about PPSD and in other current writings on assessment. This paper consults those writings in order to clarify the distinctions between the types of assessment with particular attention to their respective purposes. It then summarizes how these distinctions might inform the development of a district-wide assessment system.

DEFINITIONS

A comprehensive assessment system is comprised of three types of assessment routinely administered to all students in K-12 classrooms: summative, benchmark, and formative (Goren, 2010; North Carolina Department of Public Instruction [NCDPI], 2008). Outside of this realm are certain assessments, such as language proficiency or other diagnostic tests, given only to selected students. These assessments are not within the scope of this paper.

Stiggins, Arter, Chappuis and Chappuis (2006) consider both summative assessment and benchmark assessment to be "assessment **of** learning" employed "after learning is supposed to have occurred to determine if it did" (p. 31). Formative assessment is "assessment **for** learning," meaning that it takes place expressly to inform instruction.

The term *summative assessment* is fairly straightforward. NCDPI describes it as "a measure of achievement to provide evidence of student competence or program effectiveness" (NCDPI, 2008, p.20).

Examples of summative assessments are End of Course tests (EOC), End of Grade tests (EOG), Vocational Competency Achievement Tracking System (VoCATS) and final exams. These assessments are "high-stakes" in that their results lead to the assignment of grades, placement of students, allocation of resources, and/or the determination of federal Adequate Yearly Progress (AYP) status. Such purposes are described as *managerial*, in contrast to the *instructional* purposes described below. The intended audience—i.e., the users of high-stakes summative assessment data—includes students, teachers, principals, central administrators, board members, legislators, and taxpayers (NCDPI, 2008; Stiggins, et al., 2006).

The Council of Chief State School Officers (CCSSO) defines formative assessment as "a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes" (McManus, 2008, p. 3). This CCSSO definition has been adopted for use by NCDPI (NCDPI, 2008).

Formative assessment, therefore, should be seen as an activity rather than as one of the physical objects used to support it, e.g., student journals, ungraded tests, homework, or class work (Chappuis & Chappuis, 2008). Thus, formative assessment is considered to be embedded in the instructional process. Examples of formative assessment are descriptive feedback, teacher-student conferences, peer assessment, teacher observations, and questioning that reveals and furthers student thinking in ways that immediately influence the actions of the student and the teacher. Stiggins et al. (2006) assert that tools such as selected response, extended response, and performance assessments can be implemented formatively if they are designed to provide detailed instructional feedback and do not count towards student grades. Formative assessment serves a relatively narrow audience—students, teachers, and parents—and has the strictly *instructional* purpose of helping the teacher tailor activities to facilitate learning.

Benchmark assessment is the "middle child" of the assessment family, and seems to be the most difficult to define. Some researchers draw a slight distinction between the terms benchmark assessment and interim assessment (Herman, et al., 2010), while others consider them to be identical. This paper will use the term *benchmark assessment*, which NCDPI defines as assessment of "students periodically throughout the year or course to determine how much learning has taken place up to a particular point in time and to track progress toward meeting curriculum goals and objectives" (NCDPI, 2008, p. 14).

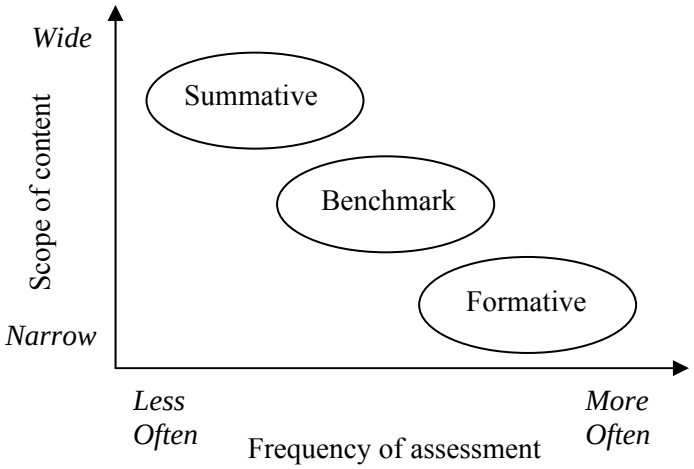
Perie et al. (2009) emphasize that benchmark assessment is characterized by data systems that aggregate results "across students, occasions, and concepts" (p. 6). It has been described by some as "early warning" or "mini-summative" in nature (Olson, 2005, p. 14). Benchmark assessment includes mid-term exams, quarterly assessments, pre-tests/post-tests, and progress monitoring.

The definitions and examples above provide a useful starting point for understanding formative, benchmark, and summative assessment. The next section examines a variety of distinctions between these types, with an emphasis on the differences between formative and benchmark assessment.

DISTINGUISHING BETWEEN FORMATIVE, BENCHMARK, AND SUMMATIVE ASSESSMENT

Two ways to distinguish between types of assessment are by frequency and scope. As shown in Figure 1, summative assessment is infrequent and covers a wide scope of content, whereas formative assessment occurs at high frequency and focuses on specific content (Perie, et al., 2009). Occupying the middle ground of this terrain is benchmark assessment, which "test[s] a slice of curriculum that is narrower than the state assessments but broader than" formative assessment (Clune & White, 2008, p. 3). Benchmark assessment occurs on an as-scheduled basis during a break in the instructional flow whereas formative assessment occurs constantly during instruction on an as-needed basis (Crane, 2008; McManus, 2008; Moss & Brookhart, 2009; NCDPI, 2008).

Figure 1
Tiers of assessment.



Adapted from Perie, et al. (2009).

To further distinguish between benchmark and formative assessment—the area in which confusion persists in many districts—other characteristics can be explored. For example, formative assessment emanates from classroom interactions and is tailored to the individual needs of students (Black & Wiliam, 1998; Chappuis & Chappuis, 2008; Marshall, 2008; Perie, et al., 2009; Stiggins, et al., 2006). Formative assessment occurs *during* the instructional unit, relying heavily on self-assessment and other forms of student involvement. This places the teacher and the student in a descriptive feedback exchange regarding specific objectives and learning dispositions, and it helps define learning gaps and map out strategies to close them (Black & Wiliam, 1998; Moss & Brookhart, 2009; NCDPI, 2008; Stiggins, et al., 2006).

Benchmark assessment does not share the characteristics described above. Instead, it is designed to summatively measure and record learning at particular points in time, when there is the

expectation—or at least the hope—that students will demonstrate mastery of material that has been taught (Chappuis & Chappuis, 2006). Usually, this means that the instruction has already been delivered. However, that would not be the case with a pre-test/post-test combination or a benchmarking system designed to track proficiencies at prescribed moments throughout a school year during breaks in the instructional flow (Downey, et al., 2009).

A benchmark assessment can be designed to serve a managerial purpose, wherein educators examine the performances of groups of students in order to monitor the effectiveness of programs and to allocate resources. Hence, benchmark assessments are typically constructed outside of the classroom with an eye towards centralizing data and emulating, in some cases, state-mandated high-stakes tests—purposes that demand certain protocols of test implementation be imposed upon the classroom teacher (Downey, et al., 2009; Perie, et al., 2009).

Benchmark assessments can be designed to serve instructional purposes by providing teachers with actionable data about the effectiveness of recently delivered instruction (Chappuis & Chappuis, 2008; Crane, 2008; Goertz, et al., 2010; Goren, 2010; Olson, 2005; Perie, et al., 2009; Stiggins & DuFour, 2009). Although such assessments may lead promptly to remediation, enrichment or other instructional responses, they do not fit the definition of formative assessment as described throughout the literature. In particular, assessments created outside the classroom typically do not provide in-depth analyses of student thinking, are not embedded in the instructional process, and do not generate the descriptive feedback characteristic of formative assessment (Black & Wiliam, 1998; Perie, et al., 2009; Stiggins & DuFour, 2009).

In addition, benchmark assessment differs from formative assessment in that it generates evaluative feedback normally presented in the form of grades, scale scores, or percentages of correct answers (Niemi, Vallone, Wang, & Griffin, 2007). Managerially purposed benchmark assessments collect and report these data primarily to serve an audience of administrators and instructional leaders, whereas the audience for instructionally purposed benchmark assessments leans more towards the classroom—i.e., students, teachers, and parents.

If the three assessment types are considered on a continuum, with formative and summative at either end, the distinctions discussed above might be seen as six dimensions along which to make comparisons, as depicted in Table 1. Because some assessment activities appear to fit into one column according to some dimensions and a different column according to other dimensions, these comparisons cannot be expected to pigeon-hole every assessment activity. Instead, the table is meant to illustrate the relationship between an assessment activity's purpose and the ways in which it is administered and interpreted.

Table 1
Viewing Assessment Types through Six Dimensions

	TYPE OF ASSESSMENT		
	*"Assessment **for** Learning"*	*"Assessment **of** Learning"*	
DIMENSION	**Formative**	**Benchmark**	**Summative**
Purpose	• Instructional	• Most designed for managerial uses • Some designed for instructional uses	• Managerial
Implementation	• Driven by moment-to-moment decisions; generated or selected by teacher; individualized	• Regulated by protocols developed in or out of the classroom; teacher-generated or externally generated	
Timing	• During instruction • High frequency	• After instruction or during a break in instructional flow • Moderate frequency	• After instruction • Low frequency
Scope	• Narrow; one or very few learning objectives at a time	• Moderate; a manageable number of objectives	• Broad; comprehensive set of objectives
Audience	• Classroom (i.e., students, teachers, and parents)	• Administration and/or • Classroom	• Public • Administration • Classroom
Feedback	• Student↔teacher • Descriptive	• System→audiences • Mostly Evaluative	• System→audiences • Evaluative

IMPLICATIONS FOR DEVELOPING A COMPREHENSIVE ASSESSMENT SYSTEM

Develop a balanced assessment system that identifies the purposes, expectations, and limitations of each type of assessment.

A plan for a district-wide assessment system should state its purposes and define all relevant terms. The plan should also identify the recipients of the resulting data and the expectations of teachers, students, and administrators. It should articulate the benefits of the assessment system and how it will help improve learning and teaching (Crane, 2008; Downey, et al., 2009; Li, et al., 2010; Niemi, et al., 2007). The system is considered balanced if each of the three types of assessment is appropriately practiced in all classrooms, with a focus placed on formative assessment (NCDPI, 2008).

Centralized data systems tend to make teachers uncomfortable about the potential use of the data (Kerr, et al., 2006). School leaders can guard against this by providing non-threatening ways for teachers to share and use their data and by adhering to the stated purposes and plans of their assessment programs. School leaders must be cognizant of all costs associated with their centrally mandated assessment programs and should assure all stakeholders that the costs are minimized and the benefits optimized (Herman, et al., 2010).

Teachers and administrators should express and promote behaviors consistent with their stated purposes (Crane, 2008). Otherwise, for example, a student falsely believing that a formative assessment activity has an evaluative purpose may employ test-taking strategies such as gravitating towards easier problems and guessing on difficult ones. Researcher Lorrie Shepard warns that while these behaviors may help the student pass a high-stakes test, they would also conceal actionable information about his/her learning (as attributed in Olson, 2005, p. 13). Thus, an assessment properly administered for its instructional value might not offer central office administrators reliable predictions of high-stakes performance or a tool with which to evaluate teachers or programs (Perie, et al., 2009).

Likewise, centrally-mandated testing protocols that ensure integrity of data across classrooms, schools, and time are not always conducive to individual student considerations and frequently can not support short-term instructional decisions (Chappuis & Chappuis, 2008). Therefore, it is not practical to expect that an assessment intended to primarily serve managerial purposes for administrators can also provide optimal *instructional* value for teachers (Christman, et al., 2009; Crane, 2008; Perie, et al., 2009).

Ensure proper alignment, and optimize the scope of content and the quantity of items.

Assessment items must be clearly aligned with the content standards and should include enough items or tasks to achieve their stated purpose without overburdening teachers or students (Gallager, 1998; Stiggins, et al., 2006). In centrally-created assessments designed for instructional purposes priority should be given to the most important objectives, especially those

that are considered key to future learning (Downey, et al., 2009; Reeves, 2000). Each of those objectives should be covered by enough items to provide a range of difficulties and cognitive demands (Herman & Baker, 2005; Niemi, et al., 2007; Olson, 2005; Reeves, 2000). While it may not be practical for centrally-created assessments to cover all objectives in a corresponding instructional time period, Downey et al. (2009) recommend that classroom-generated assessments should do so and that teachers employ pre- and post-assessments for all instructional units.

Fewer items per objective and a larger scope of objectives might be useful for benchmark assessments measuring the aggregated performance of a group of students or for high-stakes summative assessments. Such an approach would serve managerial, not instructional, purposes (Lalley & Gentile, 2009; Niemi, et al., 2007).

Design high quality assessments that offer a range of difficulties and measure a range of cognitive processes.

High-quality assessment items are written in unambiguous and non-biased language to accurately measure their targeted learning objectives. Poor items and badly constructed test forms can lead to incorrect conclusions about students' proficiencies and waste valuable class time (Gallagher, 1998; Goertz, Olah, & Riggan, 2009; Herman & Baker, 2005; Stiggins, et al., 2006).

A selected response or fill-in-the-blank item should be written to focus on only one identifiable learning objective and should not purport to measure multiple objectives. While these easily-scored items can efficiently provide valuable information about students' abilities in lower-level cognitive processes (e.g., in Bloom's Taxonomy), they offer limited opportunities to assess the higher-level processes (Gallager, 1998; Stiggins, et al., 2006). Therefore, assessments of all types—formative, benchmark, and summative—should include extended response or performance assessment items. This will produce an assessment program in which students are challenged to apply and explain key principals (Herman & Baker, 2005). Another advantage of extended response and performance assessment items is that they can measure multiple objectives if scored with analytic rubrics, which address each objective individually (Stiggins, et al., 2006).

Assessments must be consistently scored across teachers and schools if managerial purposes are to be served. Selected response and short-answer formats satisfy this need rather economically because the scoring can be done at a glance or by machines. In the case of rubric-based scoring, achieving consistency takes a considerable amount of time in both the creation of the rubrics and the training of the scorers (Arter & Chappuis, 2006). If the assessment data are to be used for strictly instructional purposes, consistency across schools becomes less important, while the need for well-designed rubrics to be used by individuals or teams of teachers working together would remain in place. In either case, the reporting system should provide results in simple, easy-to-use formats (Downey, et al., 2009).

Go beyond strategic sense-making, if the purpose is instructional.

Blanc et al. (2010) studied the ways in which elementary teachers look at student performance on common assessments[1]. They found that elementary teachers and instructional leaders engage in three different types of sense-making when sharing results at grade level meetings to help them make instructional decisions.

Strategic sense-making occurs when educators focus on the students most likely to move to the next higher level of performance and to help all students "practice" or improve test-taking strategies for the corresponding high-stakes tests. Strategic sense-making is also used to determine strengths and weaknesses across grade levels, schools, and classrooms so that resources can be more effectively allocated. Studies have shown that when meeting to share assessment results, teachers engage predominately in strategic sense-making (Black & Wiliam, 1998; Goertz, et al., 2009; Jackl & Baenen, 2010; Olah, Lawrence, & Riggan, 2010). Unfortunately, an over-emphasis on this type of conversation results in planning for rote memorization and a focus on item-driven rather than concept-driven instruction (Goertz, et al., 2009). Additionally, it fosters an unhealthy sense among students and teachers that their scores are being compared and ranked (Black & Wiliam, 1998).

Affective sense-making engages teachers and leaders to discuss "their professional agency, their beliefs about their students, their moral purpose, and their collective responsibility for students' learning" (Blanc, et al., 2010, p. 212). Affective sense-making includes supporting colleagues and finding ways to motivate students.

Reflective sense-making involves scrutinizing the content knowledge, how it is measured, and how it is best learned. During this type of conversation, teachers explore resources and consider changes to their instructional practices. They consider assessment data in light of other types of information in order to develop a complete picture of student learning. What emerges is an enhanced perspective to improve instruction and determine personal professional growth needs (Blanc, et al., 2010). Reflective sense-making depends upon a supportive environment in which teachers feel safe to question routines and assumptions (Christman, et al., 2009).

Plan district supports, instructional leadership, and staff development to enable effective use of data.

Just as high quality lesson plans and materials given to teachers do not automatically translate into improved instructional habits, neither will district-mandated assessments guarantee that teachers are learning all they can about their students' needs and reacting accordingly. Schools without good instructional leadership will not be able to respond to data, regardless of how they are presented (Shepard, 2010).

Indeed, Goertz, et al. (2010) studied the ways teachers react to common benchmark assessment

[1]A common assessment occurs when teachers collaborate to implement an assessment and/or review its results.

data in schools that provided dedicated meeting times for teacher teams, centrally-planned cycles of instruction and assessment, opportunities for professional development, and school-based instructional support personnel. The researchers made a number of observations. First, even when benchmark assessment data offered insight into elementary math students' conceptual understandings, many teachers focused on procedural proficiencies (e.g., steps in a long division algorithm) or on symptomatic behaviors (e.g., students frequently multiplying a base by its exponent).

Second, teachers with high levels of mathematical knowledge for teaching (MKT) were more likely to assess for conceptual understanding, and teachers who did so were more likely to respond by making substantive instructional changes based on the nature of their students' misunderstandings, rather than making superficial changes such as re-teaching with only slight alterations. Third, the researchers found evidence that the benchmark assessments, which themselves did not reveal much about students' thinking or problem-solving capacities, often led teachers towards classroom-based formative assessment procedures that do reveal those things (Goertz, et al., 2010).

Caveat emptor.

If the foregoing discussion of assessment types and implications provides added clarity, then it is hoped that this final implication acts as a summary and an invitation to proceed accordingly. School administrators should have a good sense of the qualities they want in assessment systems and in individual items and, if purchasing products from a vendor, should scrutinize vendors' claims about reliability, validity, and alignment with state content standards. There is widespread concern among experts that vendors rushing into a growing market are providing a large quantity of items that have not been field tested or subjected to adequate psychometric review. Furthermore, some vendors seem willing to insinuate, incorrectly, that the research supporting formative assessment also applies to benchmark assessment and test-prep products (Li, et al., 2010; Olson, 2005; Popham, 2008; Shepard, 2010).

It is imperative that school leaders understand the capabilities and costs of the systems being purchased or created by school districts or state agencies (Crane, 2008). These systems may bring convenience of accessing items, keeping records, and reporting results, but the quality of the items themselves is unlikely to be better than what textbooks have offered for years. Most importantly, those conveniences do not automatically improve student learning. On the other hand, the research is quite clear that true formative assessment practiced by teachers in the classroom does lead to substantial gains (Black & Wiliam, 1998). Given the myriad purposes, high costs, and the variety of options, districts should proceed carefully when developing and deploying comprehensive assessment systems.

REFERENCES

Arter, J., & Chappuis, J. (2006). *Creating & recognizing quality rubrics.* Portland, OR: Educational Testing Service.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139-44.

Blanc, S., Christman, J., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. (2010). Learning to learn from data: benchmarks and instructional communities. *Peabody Journal of Education*, *85*(2), 205-225. doi:10.1080/01619561003688688

Bulkley, K., Olah, L., & Blanc, S. (2010). Introduction to the special issue on benchmarks for success? Interim assessments as a strategy for educational improvement. *Peabody Journal of Education*, *85*(2), 115-124. doi:10.1080/01619561003673920

Chappuis, S., & Chappuis, J. (2008). The Best Value in Formative Assessment. *Educational Leadership*, *65*(4), 14-19.

Christman, J., Bulkley, K., Neild, R., Blanc, S., Liu, R., Mitchell, C., & Travers, E. (2009). *Making the most of interim assessments: Lessons from Philadelphia*. Philadelphia, PA: Research for Action. (ERIC Document Reproduction Service No. ED505863)

Clune, W., & White, P. (2008). *Policy effectiveness of interim assessments in Providence public schools* (WCER Working Paper No. 2008-10). Madison, WI: University of Wisconsin-Madison, Wisconsin Center for Education Research. (ERIC Document Reproduction Service No. ED503125)

Crane, E. (2008). *Interim assessments practices and avenues for state involvement.* Council of Chief State School Officers, Washington, DC.

Downey, C., Steffy, B., Poston, W., & English, F. (2009). *50 ways to close the achievement gap.* Thousand Oaks, CA: Corwin Press.

Gallager, J. (1998). *Classroom assessment for teachers.* Upper Saddle River, NJ: Merrill.

Goertz, M., Olah, L., & Riggan, M. (2009). *Can interim assessments be used for instructional change?* (CPRE Policy Briefs. RB-51). Consortium for Policy Research in Education.

Goertz, M., Olah, L., & Riggan, M. (2010). *From testing to teaching: The use of interim assessments in classroom instruction* (CPRE Research Report No. RR-65). Consortium for Policy Research in Education.

Goren, P. (2010). Interim assessments as a strategy for improvement: Easier said than done. *Peabody Journal of Education*, *85*(2), 125-129. doi: 10.1080/01619561003688688

Henderson, S., Petrosino, A., Guckenburg S., & Hamilton, S. (2008). *A second follow-up year for "Measuring how benchmark assessments affect student achievement"* (REL Technical Brief, REL Northeast and Islands 2007–No. 002). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. (ERIC Document Reproduction Service No. ED501327)

Herman, J., & Baker, E. (2005). Making benchmark testing work. *Educational Leadership*, *63*(3), 48-54

Herman, J., Osmundson, E., & Dietel, R. (2010). *Benchmark assessments for improved learning (AACC Policy Brief).* Los Angeles, CA: University of California.

Jackl, A., & Baenen, N. (2010). *Wake County Public School System (WCPSS) professional learning teams (PLTs): 2009-10 school-based policy implementation study.* Raleigh, NC: Wake County Public School System.

Kerr, K., Marsh, J., Ikemoto, G., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts. *American Journal of Education*, *112*(4), 496-520. Retrieved from EBSCO/host database

Lalley, J., & Gentile, J. (2009). Classroom assessment and grading to assure mastery. *Theory Into Practice*, *48*(1), 28-35.

Li, Y., Marion, S., Perie, M., & Gong, B. (2010). An approach for evaluating the technical quality of interim assessments. *Peabody Journal of Education*, *85*(2), 163-185. doi: 10.1080/01619561003685304

Marshall, K. (2008). Interim assessments: A user's guide. *Phi Delta Kappan*, *90*(1), 64-68.

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

Moss, C., & Brookhart S. (2009). *Advancing formative assessment in every classroom: A guide for instructional leaders.* Alexandria, VA: ASCD.

Niemi, D., Vallone, J., Wang, J., & Griffin, N. (2007). *Recommendations for building a valid benchmark assessment system: Interim report to Jackson Public Schools* (CRESST Report 723). National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles, CA. Retrieved from http://www.cse.ucla.edu/products/reports/R723.pdf

NCDPI. (2008). *Response to the framework for change: The next generation of school standards, assessments and accountability*. Raleigh, NC: North Carolina Department of Public Instruction.

Olah, L., Lawrence, N., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education*, *85*(2), 226-245. doi: 10.1080/01619561003688688

Olson, L. (2005). Benchmark assessments offer regular achievement. *Education Week*, *25*(13), 13-14

Pellegrino, J. (2004). *The evolution of educational assessment: Considering the past and imagining the future*. Princeton, NJ: Educational Testing Service.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues & Practice*, *28*(3), 5-13. doi: 10.1111/j.1745-3992.2009.00149.x

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

Reeves, D. (2000). Standards are not enough: Essential transformations for school success. *NASSP Bulletin, 84*(620), 5-19.

Sharkey, N., & Murnane, R. (2006). Tough choices in designing a formative assessment system. *American Journal of Education*, *112*(4), 572-588. Retrieved from EBSCO/host database.

Shepard, L. (2010). What the Marketplace Has Brought Us: Item-by-item teaching with little instructional insight. *Peabody Journal of Education*, *85*(2), 246-257. doi: 10.1080/01619561003708445

Stiggins, R., Arter, J., Chappuis, J., & Chappuis, S. (2006). *Classroom assessment for student learning: Doing it right—using it well*. Merrill Education/ETS college textbook series. Upper Saddle River, NJ: Pearson Education, Inc.

Stiggins, R., & DuFour, R. (2009). Maximizing the power of formative assessments. *Phi Delta Kappan*, *90*(9), 640-644.