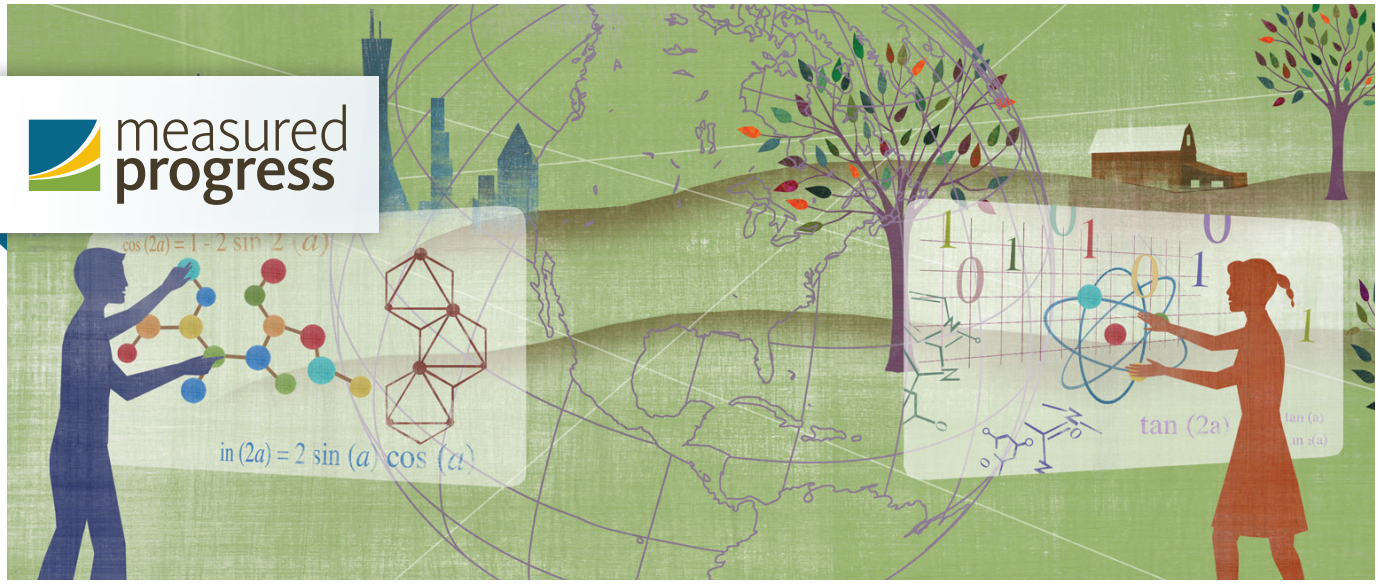

August 2012
Revised January 2013

Assessment Literacy Standards and Performance Measures for Teacher Candidates and Practicing Teachers

Prepared for the Council for the Accreditation of Educator Preparation (CAEP)

by

Stuart R. Kahl, Ph.D.

Peter Hofman, M.B.A.

Sara Bryant, M.S.

It's all about student learning. **Period.**

100 Education Way
P.O. Box 1217
Dover, New Hampshire
800.431.8901

Executive Summary

The Council for the Accreditation of Educator Preparation (CAEP) is developing new standards to accredit teacher education programs and to establish guidelines for performance measures for both teacher candidates and those already in practice. Measured Progress® was among the organizations and experts CAEP engaged to inform this effort.

This paper reflects our insights regarding the assessment knowledge and skills that should be taught in pre-service teacher education programs, as well as our recommendations related to performance measures for teacher candidates and practicing teachers.

Based upon our prior knowledge and experience, as well as research we conducted specifically for this project, we have arrived at the following conclusions and recommendations.

Conclusions:

- In many pre-service programs, the coverage of assessment literacy in course work and practica is incomplete and superficial, leaving graduates unprepared to effectively meet the demands of today's K-12 environment.
- Likewise, the most widely used performance measures cover assessment literacy superficially, rendering them incapable of gauging candidates' mastery.

Recommendations:

- Promote candidates' mastery of assessment literacy knowledge and competencies in pre-service programs by including separate course work focused on assessment, embedding assessment topics in content and methods courses, and providing real-world opportunities to enable candidates to apply what they have learned.
- Flesh out the domain of assessment literacy into objectives and learning targets to provide the specificity needed to design effective curricula, instructional materials, practica, and formative and summative performance measures.
- Evaluate programs not only in terms of the impact graduates have on student learning, but also for "inputs," such as the scope and nature of the resources and opportunities devoted to promoting assessment literacy—the course content, field experiences, and measures, all of which should be heavily performance-based.

Assessment data of all kinds are playing an expanding role in education, a trend that is here to stay. Educators must be able to effectively gather and use evidence of what students know and can do in order to foster their growth and success. The work we carried out in preparing the paper confirmed long-held perceptions that many teacher education programs and performance measures address assessment only superficially, leaving teachers unprepared to handle an essential component of effective education practice.

We hope this paper will prove useful as efforts proceed to revise the accreditation standards and, ideally, address any gaps in the coverage of these critical skills. The paper has a single focus: assessment literacy, which comprises the essential assessment-related knowledge, skills, and competencies that all teachers should be prepared to apply in K–12 classrooms. Teacher education programs should equip candidates with such literacy. To be effective, teachers must apply it and continuously build their capacity in this area. Our observations and recommendations reflect research, best practices, and our individual and collective experience.

After setting the stage with a theoretical foundation and highlights of the growing prevalence of assessment throughout K–12 education, we briefly review relevant standards and performance measures, arriving at two core conclusions:

1. Insufficient effort has been made to "unpack" existing, assessment-related standards to provide meaningful guidance for the development of pre-service program curricula, instructional materials, and practica/student teaching experiences to enable candidates to build a solid foundation in assessment literacy upon graduation and certification.
2. Although publishers of teacher certification and evaluation measures (from on-demand tests to observational protocols) have attempted to define assessment-related competencies, the resulting objectives are incomplete; the measures themselves are superficial and generally lacking performance-based requirements, which are by far the best means to gauge mastery.

In response to these conclusions and to anchor the paper in a conceptual core, we propose a high-level framework to define the domain of assessment literacy. The framework appears in Table 1 on the following page and we refer to it throughout the paper. It includes a set of assessment topics and competencies for pre-service teachers in the areas of classroom formative and summative assessment and external interim and summative (including large-scale) assessment. We also suggest how these competencies might be measured at the individual level for teacher candidates and practicing teachers, as well as at an institutional level. We caution readers that we did not design the framework as a stand-alone piece. For an accurate and complete understanding, it requires the narrative presented in Section 4. Without the narrative, readers might easily misinterpret it.

The framework's topics—knowledge and competencies—cut across all assessment types. Each topic must be further unpacked and fleshed out to create learning objectives and targets to inform teacher education programs and performance measures. Otherwise, it will be impossible to promote competence in the full domain of assessment literacy or to gauge candidates' and practitioners' ability to use assessment to help every student succeed.

Given the scope and complexities of the full domain, we expect that any effective pre- or in-service program would require at least one assessment course involving instructors with both measurement expertise and practical experience. We also expect that professors who cover content (reading educators, math educators, etc.) will need to embed in their courses certain aspects of classroom assessment, particularly those related to the formative assessment process. In Addition, measures used to gauge candidates' and practicing teachers' mastery and application of assessment literacy must be more performance-based, including not only observation but examination of artifacts.

We realize that even a high-level framework defining a domain such as assessment literacy cannot be included in an accreditation standard. However, we hope a more process-oriented standard that embodies necessary unpacking and the use of the results to inform program and performance measure design will be considered. Essentially, we strongly recommend that programs be held accountable for (1) the resources and opportunities provided to candidates to build a solid foundation in assessment by the time they graduate and (2) the scope and nature of the measures that programs use to evaluate them.

We are not the only ones to define the scope of educational assessment practices. Many organizations and individuals have done so. Indeed, numerous assessment textbooks exist, and states have accreditation standards that to varying degrees address student assessment. An excellent example of the former is Educational Assessment of Students (6th Edition) by Anthony J. Nitko and Susan Brookhart, published in February 2010. An example of the latter is the current effort by the Michigan Assessment Consortium to draft a set of assessment literacy standards.

Measured Progress *is a not-for-profit company specializing in a wide range of assessments—from effective classroom assessment practices to high-stakes, large-scale assessments for general and special student populations. We also work to build educator capacity in assessment literacy. For nearly 30 years, we have worked with state and local educators, as well as under grants or contracts with such entities as the US Department of Education, the National Assessment Governing Board, the Bureau of Indian Education, and the Bill and Melinda Gates Foundation.*

Stuart Kahl *has more than 35 years of experience in large-scale assessment. A co-founder of Measured Progress, he has led the company through a period of dramatic growth to its current position as one of the nation's foremost assessment providers. In 2010, Dr. Kahl was honored with the Association of Test Publishers Professional Contributions and Service to Testing Award for outstanding contributions to the assessment industry.*

Peter Hofman *is a 14-year veteran at Measured Progress. He has been engaged in market research, marketing/communications, strategic planning, intellectual property matters, public policy, partnerships, and various special projects. He currently serves as Vice President for Public Policy and External Relations.*

Sara Bryant *has been in education for 14 years serving as a teacher, administrator, and professional development specialist. She has been involved in assessment research and the training of teachers and administrators and is active in grant research. Her current role at Measured Progress is as Research and Development Lead for Performance Assessment.*

Table 1: The Domain of Assessment Literacy for Teachers and School Administrators

Standards	**Teachers must be able to create/select and effectively use classroom assessments for a variety of purposes.**		**Teachers and administrators must be able to select and effectively interpret and use results from external interim and summative assessments designed for a variety of purposes.**
Category of Measures	**Formative**	**Classroom Summative**	**External Interim and Summative**
Types of Measures	▪ Formative assessment evidence gathering techniques	▪ Selected-response ▪ Constructed-response ▪ Performance tasks ▪ Portfolios	▪ District benchmark ▪ Diagnostic ▪ General achievement ▪ Adaptive ▪ State accountability
Quality of Measures	▪ Unpacking standards ▪ Depth of knowledge ▪ Quality of evidence regarding learning targets	▪ Good and bad items/tasks ▪ Reliability and validity · Test length · Domain representation (See "Alignment")	▪ Match to purpose ▪ Universal Design ▪ Item quality in banks and tests ▪ Item selection criteria ▪ Alignment · Categorical concurrence · Depth of knowledge · Range of knowledge · Balance of representation ▪ Technical characteristics (reliability, validity)
Results and Their Use	▪ Quality and use of feedback ▪ Use of data to inform instruction	▪ Scores vs. grades ▪ Effective and detrimental grading practices	▪ Reporting statistics · Scaled scores · Percentile ranks · Performance levels ▪ Subgroup/subtest results ▪ "Growth" and longitudinal data ▪ Comparability issues

Table of Contents

List of Tables

1. Introduction

The Council for Accreditation of Educator Preparation (CAEP) has initiated a series of verbal and written exchanges among measurement experts to inform CAEP's efforts to establish new standards for the accreditation of teacher education programs. Through this effort, CAEP also seeks to provide guidelines for performance measures, both for teacher candidates and practicing educators. These efforts are spurred by concerns that the existing standards and performance measures do not adequately address the knowledge and skills teachers need in order to use assessment effectively to evaluate and improve student learning.

We at Measured Progress appreciate the privilege of participating in this critical work. We have prepared this paper to provide information we believe will be useful to CAEP and its assembled experts in preparing the new standards and guidelines.

Problem Statement

Given the importance of teacher assessment literacy, colleges of education and K–12 schools need sufficient guidance to build this capacity for both pre-service and practicing teachers. Although most teacher preparation standards refer to assessment-related competencies for teachers, the level of specificity tends to be inadequate. This is one reason why assessment literacy is most likely absent or short-changed in teacher education program curricula, instruction, and practica. Consequently, we believe that many, if not most, student teachers and practicing teachers are not prepared to use assessment effectively to promote student learning, analyze data, and make decisions from such data; that is, to be effective consumers and users of assessments.

Similarly, measures of teacher candidates and practicing teachers' assessment literacy tend to be deficient in this area. Licensure tests, portfolio exhibitions, and other measures typically don't reveal what teachers know and are able to do with assessment literacy concepts and skills.

Some sets of standards are more detailed than others, such as those from the International Reading Association and the Interstate Teacher Assessment and Support Consortium (InTASC), and competencies cited by existing performance measure programs, notably The Principles of Learning and Teaching in The Praxis Series™ from the Educational Testing Service. However, the problems cited above stem from a significant shortcoming in how the standards are used. Implementing education standards requires a process commonly referred to as "unpacking the Standards." The unpacking process typically involves breaking the standards down into user-friendly learning targets that map against a given learning progression of skills and processes. The more specific the standards, the more easily and accurately they can be unpacked. Without such unpacking, educators are at a loss in establishing content curricula, planning instruction, and building assessments that promote mastery of the standards. This is as true for pre-service programs as it is for K–12 programs.

Unpacking takes place at different levels in K–12 education. For example, the Smarter Balanced Assessment Consortium used a multi-step process typical of state assessment programs to build the bridges from the Common Core State Standards (CCSS) to specific assessment instruments. The steps included developing content specifications (with assessment targets), item and task specifications (with samples), and test specifications (with a blueprint) (Measured Progress & ETS Collaborative, 2012, p. 8-10). At another level, local education agencies typically develop curriculum frameworks from standards to provide the detail needed to create or select curricula, as well as instructional materials and assessments. In addition, teachers undertake in-depth unpacking to identify underlying constructs within each standard as a foundation for planning instruction and assessment. At first, unpacking might even require more time and effort than planning.

While unpacking is a standard practice in K–12 education, it is not as prevalent in pre-service programs, especially concerning assessment literacy. While individual programs might unpack the assessment domain, we believe the practice is far from widespread. We have found that well-known performance measures measures for candidates or practicing teachers attempt to define the domain of assessment, but in each case both the unpacking and measures are incomplete.

Purpose and Structure of the Paper

The purpose of this white paper is to recommend assessment literacy standards for teacher educators and to specify how these standards could be measured.

To put some context around what informed our recommendations, this paper also includes:

- background information—including brief assessment theory, research, and future trends;
- an overview and synthesis of the extent to which existing standards and measures include assessment literacy components.

First, we provide some examples of assessment literature that touch on assessment literacy in teacher education programs and as part of K–12 professional development initiatives. We also devote a section to how the current landscape and future trends frame the importance of teachers mastering assessment knowledge, skills, and practices and applying them in their classrooms to promote student learning.

In the next section, we compare several sets of standards from reputable educational organizations with the domain presented in Table 1. These standards came from:

1. AdvancED®
2. American Federation of Teachers, National Council on Measurement in Education, National Education Association (AFT/NCME/NEA)
3. International Association for K-12 Online Learning (iNACOL)
4. CCSSO's The Interstate Teacher Assessment and Support Consortium
5. International Reading Association (IRA)
6. National Board for Professional Teaching Standards (NBPTS)
7. National Council for Accreditation of Teacher Education (NCATE)
8. National Council of Teachers of Mathematics (NCTM)

Next, we briefly examine three assessments to determine to what extent assessment literacy domains are measured; the Praxis II (Educational Testing Service [ETS], 2012), the National Evaluation Series™ (Pearson Education, Inc. 2011) and the National Board Certification assessment (Pearson, 2012). We offer brief summaries of our impressions of these assessments based on limited data retrieved from their respective websites. We also discuss "The Framework for Teaching Evaluation Instrument: 2011 Edition" by Charlotte Danielson—another approach to examine teacher practice. Although this tool is not used for certifying teachers, it is worth noting because it relies on observational data across many assessment literacy domains.

In the final section, we offer recommendations of assessment literacy areas that can be used to develop specific learning objectives and targets. We also comment on how these targets might be measured in teacher education programs and as part of teacher certification and evaluation.

Background and Context

Broad Definition of Assessment Literacy

Assessment literacy encompasses the knowledge and skills educators need to

1. Identify, select, or create assessments optimally designed for various purposes, such as
   a. Accountability
   b. Instructional program evaluation
   c. Student growth monitoring and/or promotion
   d. Diagnosis of specific student needs (learning gaps)
2. Analyze, evaluate, and use the quantitative and qualitative evidence generated by external summative and interim assessments, classroom summative assessments, and instructionally embedded formative assessment practices to make appropriate decisions to improve programs and specific instructional approaches to advance student learning. Appropriate decisions depend upon a good understanding of test quality considerations and comparability issues.

This definition serves as the foundation for the assessment literacy domain framework appearing in Table 1 and described in detail in Section 4.

In general, teachers and administrators need comparable levels of expertise in assessment literacy. While teachers must practice the knowledge and skills daily in their classrooms, administrators must (1) provide the appropriate opportunities for professional

development and ongoing collaboration to sustain this competency, (2) practice it at the school or district level, and (3) evaluate teachers' assessment practice for both formative and summative purposes.

Theory and Research on the Impact of Assessment Literacy

Understanding what students know and can do is essential to effective teaching. Assessment practices, both formative and summative, rely on a core set of skills and specialized knowledge, which, when applied, are a significant component of teacher effectiveness. An abundance of literature exists regarding the impact of teachers assessment literacy on student motivation and achievement. The education field has long agreed that teachers need to learn about and implement sound assessment practices. Several landmark reviews of assessment literacy components have substantiated why the knowledge, skills, and practices embedded in assessment literacy are such fundamental and powerful components of effective teaching.

For example, Paul Black and Dylan Wiliam (1998a) examined 250 studies on assessment-related concepts and found that when teachers practice high-quality formative assessment, student achievement increases by effect sizes of .4-.7 standard deviations. Black et al. (2003) report, "Such effect sizes are among the largest ever reported for sustained educational interventions." Most notably, Black and Wiliam found that struggling students (i.e., students identified with learning disabilities, students who lack motivation) benefit most from effective formative assessment. Thus, the implementation of such assessment practices has been shown to improve outcomes for all students and essentially close the achievement gap. (See Black & Wiliam, 1998a for the complete review and 1998b for the summary of the review.)

In another landmark effort, the Committee on the Foundations of Assessment, with editors Pelligrino, Chudowsky, and Glaser, produced "Knowing What Students Know," 2001. The book highlights aspects of assessment design and practices, as well as how teachers might use assessment. Each section of the book is tied to theoretical underpinnings that support such recommendations.

In another example, evidence of the positive impacts and implications of assessment was presented in Congressional testimony regarding the proven benefits of performance assessment. These include promotion of higher-order thinking and other so-called 21st century skills, as well as increased student motivation (Wood et al., 2007).

Educational assessment experts have also reviewed and drawn on research of important assessment literacy concepts, recommending those they believe should be addressed in teacher education programs and professional development settings. Rick Stiggins, most notably, has written extensively on assessment literacy, balanced assessment systems, and the importance of sound assessment development, use, and communication (Stiggins, 2007). W. James Popham has written numerous books and articles over the last 40 years, arguing for sound assessment practices (see Popham, 2003, as an example). These references are particularly noteworthy because both Popham and Stiggins have worked directly and extensively with educators, making their contributions to the body of knowledge both relevant and practical.

Several book chapters and articles regarding what teacher candidates should know and be able to do are also worth noting. For example, a book edited by John Bransford and Linda Darling-Hammond, "Preparing Teachers for a Changing World," includes a chapter on formative and summative assessment (Shepard, et al., 2005). Another example is "Educational Assessment of Students" (6th Edition) by Anthony J. Nitko and Susan M. Brookhart (Feb 26, 2010), a dense 500+ book dedicated to assessment knowledge, skills, and practices for educators.

The importance of assessment as an essential component of effective teaching is not a new concept, as borne out by the fact that most of the sources cited here are at least several years old. Yet, it appears that many pre-service programs have still not adequately addressed this area. As described in the next section, current demands and future trends will make coverage of assessment literacy in teacher education programs even more critical to teacher effectiveness—and student success.

The Role of Assessment: Current Landscape and Future Trends

The role of "data" (broadly defined) in public education has never been greater. This reality is driven by the growing importance of research-based practices, data-based decision making, and outcomes-oriented

accountability. In the wake of NCLB and its high-stakes accountability provisions, the volume of annual statewide testing has doubled over the last decade. The use of interim and other tests has grown exponentially. It is a rare district indeed that has not expanded its use of assessment for both traditional purposes and new ones, such as Response to Intervention, which relies heavily on "progress monitoring" via assessment.

From a purely logical perspective, the growth of assessment makes sense: sound evidence can inform efforts to help students, teachers, schools, districts, and even states improve. Without it, educators are shooting in the dark. Unfortunately, not all data provide sound evidence for the decisions they are supposed to inform, and all too frequently educators misinterpret and misuse the data they have—even "good" data. The reason: many educators are not assessment literate.

Current trends in education reform point to an increased use of data, along with growing demands for educators at all levels (Pre-K–20) to be assessment literate. Here are a few examples to illustrate this point.

1. Despite all the issues raised about NCLB, the role of annual state testing not only appears secure, but rising in importance as student outcomes drive multiple factors related to public education: state compliance with the AYP provisions of NCLB or with commitments made under USED waivers as well as Race to the Top awards; teacher and administrator evaluations for accountability and staff development purposes; and teacher education program accountability evaluation.
2. The widespread adoption of the Common Core State Standards (CCSS) will greatly increase the complexity and performance orientation of assessments built to measure student progress relative to standards. The new emphasis on higher-order skills and even new content will require a greater understanding of multiple aspects of assessment that will be new for many educators. The five consortia developing new assessments for the CCSS will usher in novel assessment instruments—such as Technology-Enhanced Items (TEIs)—laying the foundation for future innovation, perhaps involving sophisticated simulations and gaming. The new instruments will require educators to be assessment literate, so they can fully understand and apply results.
3. The increasing use of technology in K–12 education—from an abundance of online assessment resources to virtual and hybrid teaching models and adaptive instructional materials with embedded assessments—will require careful educator scrutiny to develop new means of gathering evidence of student learning, understand the new assessment content, and interpret and appropriately use the results.

The bottom line is that assessment will always play an essential role in education. Proven traditional forms of assessment will undoubtedly persist even as new standards and measures are adopted. It will take assessment literate educators to both effectively use existing measures and adapt to and take advantage of new tools in order to consistently maximize student success in the broadest sense.

2. Coverage of Assessment Literacy in Major Education Standards

Many organizations have developed standards for teacher education and in-service programs. The spectrum ranges from content-specific standards for areas like math and English language arts to standards designed for specific types of educators, such as para-professionals and literacy specialists. The depth and breadth of standards also vary from organization to organization. Some standards, perhaps the largest group, barely mention assessment, if at all. Others only consist of broad principles that could be looked at as philosophical or visionary statements that are too general to inform a teacher education program and performance measures. In selected areas, standards exist that reflect some unpacking and begin to be specific enough to inform curriculum development and instructional planning. We found no standards that encompassed all the high-level topics presented in Table 1.

For the purposes of this paper, we have used the assessment literacy domain framework (Table 1) to tease out the coverage of assessment literacy in standards from a wide variety of reputable educational organizations. The framework is organized by assessment literacy components we believe are central to an educator becoming assessment literate. Section 4 describes the framework for the domain of assessment literacy. It explains some of the entries, providing examples of current misunderstandings and their implications.

Table 2 presented on the following page attempts to compare our proposed framework with existing standards. The codes used in Table 2 are meant to capture the extent to which existing standards incorporate the assessment literacy components, "Type of Measure," "Quality of Measure," and "Results and Use" by the three different types of assessment (formative, classroom summative, and external interim/summative). A "B" in a cell indicates the organization broadly includes the respective assessment literacy component in its standards. An "S" indicates more specificity is represented. An empty cell indicates there was no evidence of the assessment literacy component.

We do not intend for the table to reflect an evaluation of the quality of these standards. Rather, it represents a quick snapshot to illustrate how specifically several sets of standards address the topics identified within the domain. We used the following guiding question to assign a "B" and "S" to each set of standards: "When thinking about the assessment literacy component _____, are the standards too broad or are they specific enough to start to inform the development of curricula, instructional activities, and performance measures for teacher candidates or practicing teachers?"

Based on our review, we can state that assessment literacy components are visible in all eight of the standards surveyed. This visibility is, however, a bit deceiving because in reference to the three assessment literacy components in our framework, there is limited evidence of any kind of specificity within the standards.

For example, Table 2 shows there are "Broad" InTASC standards for "Quality of Measures" for classroom summative assessment, yet InTASC doesn't reference external interim and summative assessment: "The teacher understands the differences between formative and summative applications of assessment and knows how and when to use each." (p. 15).

In contrast, the IRA reaches a deeper level of specificity within its "Quality of Measures" standards for all types of assessment. For example, in Element 3.2 for middle level teachers, a formative assessment indicator reads, "Interpret and use assessment data to analyze individual, group, and classroom performance and progress." (IRA, 2011). Although this indicator has room to offer more specificity, it at least goes beyond the general term "analyze data" to reference the importance of analysis at the individual, group, and classroom levels.

To illustrate the range of specificity with which different organizations' standards address components of assessment literacy, we have provided examples from four sets of standards in Table 3, which follows. The side-by-side comparison should clearly demonstrate the difference between broader and more specific standards. Note that we did not find any NCATE standards related to assessment that we thought warranted classification as specific.

Given our analysis of these standards we offer several preliminary conclusions about the specificity of

Table 2: The Coverage of Assessment Literacy in Major Education Standards

Assessment Literacy Component	Organizations' Standards	Formative	Classroom Summative	External Interim and Summative
Type of Measure *Standards differentiate between types of assessment measures.*	AdvancED[1]	—	—	—
	AFT/NCME/NEA[2]	B	S	S
	iNACOL[3]	B	B	—
	InTASC[4]	S	B	B
	IRA[5]	S	S	S
	NBPTS[6]	B	—	—
	NCATE[7]	B	B	—
	NCTM[8]	B	B	—
Quality of Measures *Standards reference how to develop quality measures and/or judge the quality of measures.*	AdvancED	—	—	—
	AFT/NCME/NEA	B	S	B
	iNACOL	B	B	—
	InTASC	S	B	—
	IRA	S	S	S
	NBPTS	B	—	—
	NCATE	B	B	—
	NCTM	—	—	—
Results and Use *Standards reference how to appropriately use assessment results.*	AdvancED	—	S	—
	AFT/NCME/NEA	S	S	S
	iNACOL	S	B	—
	InTASC	S	B	—
	IRA	B	B	B
	NBPTS	S	B	—
	NCATE	B	B	B
	NCTM	B	—	—

KEY	
B	Broadly includes the assessment literacy component in their standards
S	More specificity is represented
—	No evidence of the assessment literacy component

[1]AdvancED®, 2011

[2]American Federation of Teachers, National Council on Measurement in Education, National Education Association (AFT/NCME/NEA), 1990

[3]International Association for K-12 Online Learning (iNACOL), 2011

[4]CCSSO's The Interstate Teacher Assessment and Support Consortium, 2011

[5]International Reading Association (IRA), 2010

[6]National Board for Professional Teaching Standards (NBPTS), 2012

[7]National Council for Accreditation of Teacher Education (NCATE), 2008, p. 19

[8]National Council of Teachers of Mathematics (NCTM), 2011

existing standards with regard to assessment literacy and the feasibility of using the standards to develop sound assessment-related curricula, instruction, and assessments.

1. Most standards are too broad to inform the development of high-quality curricula, instruction, and assessment—leaving too much potential for shortchanging assessment topics by some institutions and schools.
2. There tends to be an emphasis on broad formative assessment constructs.
3. Standards are lacking for knowledge about interim assessment and large-scale, summative assessment. (While some people may think an understanding of these assessments is more appropriate for administrators as we will note later, teachers and teacher committees are often assigned the task of interpreting and using results from these "external" measures as well.)
4. There is a lack of differentiation between formative and classroom summative assessment.

As part of our research, we reviewed some other teaching standards. Rhode Island's include several competencies around assessment, reflecting some unpacking (Rhode Island Department of Elementary and Secondary Education, 2011). Most of the competencies are specific, but the scope is narrow, only addressing classroom assessment and focusing somewhat on elements of formative assessment practice.

In contrast, no official state standards in Massachusetts (Massachusetts Department of Elementary and Secondary Education, 2009) or Michigan (Pearson Education, Inc., 2011b) address assessment, and the states' teacher certification tests do not appear to cover it. On the other hand, the Michigan Assessment Consortium, a not-for-profit organization comprised of education agencies and their staffs from across the state, released the latest version of its comprehensive set of *Assessment Literacy Standards* in September. The document contains distinct sets of standards for teachers, administrators, policy makers, and students (and their parents).

We also reviewed Draft #5 of the *Classroom Assessment Standards: Sound Assessment Practices for PK–12 Teachers* developed by the Joint Committee on Standards for Educational Evaluation. While these standards offer many specifics, as the name implies, they only address classroom assessment and don't cover external interim and summative assessment.

Table 3: Examples of Broad and Specific Standards Related to the Assessment Literacy Domain

Organization	Broad Standard	More Specific Standard
NCATE	Candidates in advanced programs for teachers have a thorough understanding of assessment. They analyze student, classroom, and school performance data and make data-driven decisions about strategies for teaching and learning so that all students learn. (p. 19)	N/A
InTASC	6(c) The teacher works independently and collaboratively to examine test and other performance data to understand each learner's progress and to guide planning. (p.15)	6(d) The teacher engages learners in understanding and identifying quality work and provides them with effective descriptive feedback to guide their progress toward that work. (p. 15)
NBPTS	In the Level 4 performance, the teacher thoughtfully engages in insightful reflection through critical analyses and evaluation of classroom practices to make thoughtful suggestions for future instruction. (p. 2-1)	The Level 4 performance provides clear, consistent, and convincing evidence: that the teacher thoughtfully formulates purposeful, short-term and long-term, data-driven instructional goals that are firmly based on local, state, and/or national standards and curricula. (p. 2-1)
IRA	Explain district and state assessment frameworks, proficiency standards, and student benchmarks.	Recognize the basic technical adequacy of assessments (e.g., reliability, content, and construct validity).
iNACOL		Standard H: The online teacher is able to apply authentic assessments as part of the evaluation process, assess student knowledge in a forum beyond traditional assessments, and monitor academic integrity with assessments. P. 12

3. Summary of Assessment Literacy in Current Measures for Evaluating Teacher Candidates and Practicing Teachers

This section examines measures used to evaluate what teacher candidates and teachers know about assessment literacy and what they are able to implement in classrooms. We included the following measures in our review:

- The Praxis Series™ from ETS, which is a popular assessment used for evaluating and credentialing teacher candidates. The Praxis Series is designed for college students entering teacher education programs and teacher candidates needing certification to become teachers. The Praxis II® Subject Assessments are designed to evaluate teacher candidates' content and professional knowledge. From the 120 Praxis II® tests, most assessment knowledge is measured by the Principles of Learning and Teaching (PLT) Tests at the elementary, middle, and secondary levels. It's worth noting that the 2011 edition of the PLT tests devotes between 11 to 15 percent of the items to assessment, all of which are multiple-choice (ETS, 2011).
- The National Evaluation Series™, Assessment of Professional Knowledge—Elementary from Pearson, is also a popular credentialing assessment for teachers. We referred to the Test Framework for this assessment. It appears that roughly 10 percent of the test score addresses assessment literacy content and skills (Pearson Education, Inc., 2011a). It should be noted that states such as Massachusetts and Michigan also use tests published by Pearson Education, Inc. and individualized for each state. These NES exams are content specific and in these specific states, measure only teacher candidates' content knowledge (e.g., science and math content). Based on test objectives, assessment literacy is not present. (see Massachusetts Department of Elementary and Secondary Education, 2009 and Pearson Education, Inc, 2011b).
- The National Board, used for awarding National Board Certification® from the National Board for Professional Teaching Standards is designed to award an advanced teaching credential to practicing teachers. Teachers are assessed on their completion of a portfolio and six "assessment exercises." Using the NBPTS standards (summarized in Table 2), scoring guides are used to judge assessment-related standards such as using appropriate assessment methods and making instructional adjustments.
- The Framework for Teaching Evaluation Instrument—2011 Edition by Charlotte Danielson. This instrument reflects prior research and development by this education expert and is designed to inform teacher observation as an input to formative and summative evaluation of teachers. It is distinct from the three preceding instruments because it is not used for credentialing.

We note that the Praxis and National Evaluation Series dominate the teacher credentialing market and that the National Board has garnered substantial respect within the profession.

For each of these measures we were only able to review publicly available information. Understandably, while the assessment frameworks were readily available, the publishers have made public only a small sampling of items. Charlotte Danielson's criteria and rubrics are all published. We compared this information with the assessment literacy domain framework presented in Table 1 (and described in detail in Section 4). Because of the very limited availability of the actual test content, we were unable to dig deeply to fully judge the coverage of assessment literacy. We compare the three assessment programs in Table 4 on the following page. Instead of labeling cells with a "Broad" or "Specific" indicator (as we did with standards), we simply placed an "X" in each cell to represent any level of evidence of measurement.

Our review of the above-cited measures was particularly revealing. In general, these entities attempt to define the assessment literacy domain, in some cases by unpacking assessment-related standards, typically breaking them down to "competencies." In some cases, the competencies were more specific than what we reviewed in the standards. Danielson's framework even presented rubrics to gauge levels of mastery of the identified competencies. In other cases, the unpacking was incomplete—only removing the bubble wrap: competencies appeared to be simply a restatement of the overarching standard. In no case did the measures come close to covering even our high-level framework for the domain.

The measures leaned heavily toward classroom assessment, particularly components of the formative assessment process. While we acknowledge the documented power of formative assessment to promote student learning, other forms of assessment—both classroom summative and external assessment—play important roles in the teaching/learning process. Their inadequate coverage leaves gaping holes in skills critical for teacher effectiveness. For example, by focusing on formative assessment or at the test—rather than the item—level, these competencies miss essential components of assessment literacy, from the multiple types of alignment and considerations related to item quality to issues surrounding comparability.

In many cases, the competencies (knowledge and skills) identified reflect the performance/process nature of assessment, typically using verbs like "demonstrate" and "apply." However, in many other cases the competencies call for understanding reflected

Table 4: The Coverage of Assessment Literacy in National Measures for Evaluating Teacher Candidates and Practicing Teachers

Assessment Literacy Component	Organizations' Standards	Formative	Classroom Summative	External Interim and Summative
Type of Measure *Standards differentiate between types of assessment measures.*	National Board[9]	X	X	
	NES[10]	X	X	
	PRAXIS[11]	X	X	X
Quality of Measures *Standards reference how to develop high quality measures and/or judge the quality of measures.*	National Board			
	NES	X	X	X
	PRAXIS	X	X	
Results and Use *Standards reference how to appropriately use assessment results.*	National Board	X	X	
	NES	X	X	
	PRAXIS	X	X	X

KEY	
X	Evidence of any coverage

[9]National Board for Professional Teaching Standards®: Early and Middle Childhood Literacy: Reading-Language Arts—Scoring Guide for Candidates

[10]NESR® National Evaluation Series™–Assessment of Professional Knowledge—Elementary: Test Framework, 2011. Pearson's, NES provides states with multiple testing options for entry level teachers.

[11]Praxix Series: The Praxis serves a significant portion of the market for teacher credential exams.

in the test-takers' ability to define or explain terms and concepts, which we think promotes a superficial and simplistic view of important assessment literacy topics.

Despite the performance orientation of many of the defined assessment competencies, the measures themselves offer at best a mixed bag. The Praxis Series and the NES rely on multiple-choice items. (We do not know if the writing component case study in the NES touches on assessment.) Most of the samples we found would not tax a candidate's ability to apply the knowledge and skill gained in pre-service programs. Indeed, they tend to promote shallow knowledge and simplistic awareness of concepts, many of which have inherent complexities. In one case, we were shocked by the bias toward multiple-choice items embedded in the item, which, it could be argued, took a position contrary to sound measurement principles. Moreover, the test frameworks have so few items that adequate coverage of even their limited definition of the domain was impossible. Available documentation does not indicate domain coverage.

On the other hand, the National Board assessment is, despite its narrow scope, more performance-oriented. It consists of a portfolio with multiple artifacts (e.g. lessons, videos, reflections) and an "assessment experience" (i.e., constructed-response items). We should also note that the Teacher Assessment Portfolio Consortium (TPAC), a collaboration of 25 states, is implementing the Teacher Performance Assessment (TPA). The TPA examines factors and uses, process, measures, and rubrics similar to the National Board assessment, except that it is intended for teacher candidates rather than practicing teachers. Pilot testing is complete and initial implementation is scheduled for this fall, with revisions anticipated in the spring of 2013. While we support its performance orientation, we caution that if its scope mimics that of the National Board assessment, it will overlook portions of the assessment literacy domain we think are essential.

The Framework for Teaching Evaluation Instrument is designed to address an increasingly important component of measuring teacher effectiveness: observation. The framework does a reasonable job of covering the formative assessment process, but barely addresses classroom summative assessment and such key factors as technical quality (e.g., the different types of alignment). It also doesn't cover external interim and summative assessment. The Framework presents a comprehensive guide to observation. What is less clear is the extent to which the process includes reviewing teachers' artifacts—lesson plans, tests, descriptive feedback, and adjustments to instruction based upon evidence of student learning.

Based on our review, we can draw a few general conclusions about how existing performance measures for teacher candidates and teachers address our proposed assessment literacy domain framework.

1. Although the measures reflect some effort to define the competencies within the domain, even in the best of cases, the coverage is incomplete. In general, the measures tend to more completely address the assessment strategies in the formative assessment process, which can dramatically boost student learning. Nevertheless, we think that the overlooked topics—and even components—of assessment literacy are critical. These topics range from important considerations in item/test development/selection (such as alignment, reliability, and validity) to important factors (such as comparability) in interpreting and using the results from different student assessments.
2. The shortcomings of the incomplete definition of the domain are magnified by the scope and nature of the two most popular performance measures, The Praxis Series and NES. The fact that they rely almost exclusively on multiple-choice items to measure pre-service candidates' mastery of a body of knowledge and skills that will be manifested through performance over-simplifies and in many cases superficially covers the domain, calling for recall rather than application.
3. By being performance based, The National Board assessment and TPA are headed in the right direction, but their scopes overlook essential components of assessment literacy.

4. Recommendations for Assessment Literacy Standards and Performance Measures

Overview

On the basis of the information provided in the preceding sections, we can draw three general conclusions:

1. The need for educators to be assessment literate has never been greater and, given current trends in education reform, its importance will only continue to grow.
2. Existing standards—whether for graduation, licensure, professional certification, or accreditation—vary in the extent to which they address the body of knowledge and skills encompassed by assessment literacy. In general, they do not provide sufficient specificity to guide pre-service curricula and instruction to promote assessment literacy or to inform the design of performance measures to evaluate the knowledge and skills of teacher candidates and the effectiveness of practicing teachers.
3. Following the pattern we observed in the standards, existing teaching performance measures do not adequately address the breadth and depth of assessment literacy; even those that include multiple sources of evidence typically overlook artifacts that often best illustrate mastery. We also argue that strictly multiple-choice assessments, like the Praxis II and NES test, do not adequately measure the complete assessment literacy domain as proposed in our framework, let alone the general standards that already exist.

In this section, we propose a framework that we hope will drive the development of appropriate teacher education program accreditation standards on assessment literacy, as well as pre-service and in-service performance measures. The framework consists of the domain of assessment literacy for teachers and administrators, which we depict in the form of a matrix in Table 1, presented again on the following page.

The matrix has a straightforward structure that immediately provides far more detail than most existing standards. The top row unpacks a single standard that all groups seem to espouse: creating/selecting and using assessments for a variety of purposes. Note that we have specified both creating and selecting assessments; the same principles apply to each. We have witnessed an explosion of third-party assessment content available to districts, schools, and individual teachers—from adaptive and fixed-form tests to item banks. Much of this content has not been vetted, especially that which is available online. Even assessment resources developed by reputable entities might be inappropriate for a particular purpose. Assessment literate teachers know they have to evaluate third-party content with the same considerations they would use in creating their own. As the range and volume of these resources continue to expand, the ability of teachers to evaluate them will similarly grow in importance.

We have broken this standard into two parts, one focused on classroom assessment and the other on tests from "external" sources. We have divided classroom assessment into two distinct areas: formative and summative. As we will describe below, we adhere to the definition of formative assessment researched and espoused by Black and Wiliam, 1998. We recognize that educators can, in some circumstances, use the results of summative assessments in a formative manner, but that does not generally qualify them as formative assessment. The distinction is sufficiently great to warrant separate consideration of the knowledge and skills required to effectively use each.

The three functional rows in the matrix address the essential components needed to achieve the overall standard. Types of Measures and Quality of Measures relate to creating/selecting assessments. Results and Their Use encompasses the knowledge and skills needed to fulfill the ultimate purpose of each assessment.

To design the matrix as efficiently as possible, we assumed a horizontal flow of topics in both directions, rather than trying to make each column independent of one another.

For example, a solid grasp of the different types of measures listed in the Classroom Summative column will enable a teacher to appropriately use the measures in the evidence-gathering step of formative

Table 1 (repeat): The Domain of Assessment Literacy for Teachers and School Administrators

Standards	**Teachers must be able to create/select and effectively use classroom assessments for a variety of purposes.**		**Teachers and administrators must be able to select and effectively interpret and use results from external interim and summative assessments designed for a variety of purposes.**
Category of Measures	**Formative**	**Classroom Summative**	**External Interim and Summative**
Types of Measures	▪ Formative assessment evidence gathering techniques	▪ Selected-response ▪ Constructed-response ▪ Performance tasks ▪ Portfolios	▪ District benchmark ▪ Diagnostic ▪ General achievement ▪ Adaptive ▪ State accountability

Quality of Measures	▪ Unpacking standards
	▪ Depth of knowledge
	▪ Quality of evidence regarding learning targets
		▪ Good and bad items/tasks
		▪ Reliability and validity
		  ◦ Test length
		  ◦ Domain representation (See "Alignment")
			▪ Match to purpose
			▪ Universal Design
			▪ Item quality in banks and tests
			▪ Item selection criteria
			▪ Alignment
			  ◦ Categorical Concurrence
			  ◦ Depth of knowledge
			  ◦ Range of knowledge
			  ◦ Balance of representation
			▪ Technical characteristics (reliability, validity)

Results and Their Use	▪ Quality and use of feedback
	▪ Use of data to inform instruction
		▪ Scores vs. grades
		▪ Effective and detrimental grading practices
			▪ Reporting statistics
			  ◦ Scaled scores
			  ◦ Percentile ranks
			  ◦ Performance levels
			▪ Subgroup/subtest results
			▪ "Growth" and longitudinal data
			▪ Comparability issues

assessment or in contributing to or appropriately using external tests. As another example, which we refer to more specifically below, sound knowledge of the key considerations in alignment and technical characteristics that are cited in the External Interim and Summative column will provide valuable insights to teachers as they create/select classroom assessments. And across the board, matching an assessment with its purpose is essential. Yet we find from the classroom teacher to state education agencies a prevailing tendency to assume a single test can serve a multitude of purposes. It can't.

We expect questions will arise about how Universal Design, accessibility, and assessing students with special needs fit into the assessment literacy framework of Table 1. For more than a decade, educators and researchers have applied the principles of Universal Design to student assessment. Originally developed by architects, Universal Design involves designing and developing products to function appropriately for a broad spectrum of people while also supporting extensions to meet specific access needs. In the case of assessment, this means designing testing instruments, items, and tasks with the spectrum of users in mind up front, rather than retrofitting them after the fact. The principles underlying Universal Design can be applied to state, district, school, and classroom assessments.

In general, the principles and research-based practices underlying the matrix apply across the board to all students. We have included Universal Design in the External Interim and Summative column because this set of principles has become a standard feature of all statewide and many commercial assessment content development efforts. Teachers should certainly consider the principles as they create classroom assessments.

We must note that the table entries are, in many cases still broad topics. Each could be "unpacked" into multiple objectives or learning targets characterizing what teachers should know and/or be able to do. Thus, we recommend that the table serve as a starting point for a panel of experts to use in developing a coherent and comprehensive set of objectives/learning targets. Even at this level, we think that the matrix fills an important gap by starting to unpack the broadest, generally accepted standard, breaking it down into topics that can be fleshed out into specific objectives/learning targets. The result will be valuable guidance for designing course curricula and field experience to build assessment literacy and for developing performance measures for teacher candidates and practicing teachers. We have no doubt that promoting mastery of the body of knowledge and skills encompassed by assessment literacy will greatly enhance how effectively—and efficiently—teachers use assessment to promote student learning.

Our many interactions over the years with teachers and administrators, through our professional development work and in conjunction with our state testing programs, have suggested to us that significant gaps exist in practitioners' understanding of the topics identified in the table. While this paper focuses on the assessment literacy of teacher education students and practicing teachers, we include administrators in our scope for three reasons. First, the majority of school administrators are former teachers. Thus, the domain of assessment literacy is the same for them as it is for teachers. Second, while one might think that understanding and using external assessments are more relevant to administrators, typically administrators pass the results on to teachers or teacher teams to interpret or, as is too often the case, to over- or mis-interpret. Third, teacher teams often create district-level tests.

The remainder of this section elaborates on the contents of Table 1, addressing the cells in order by column. Throughout the discussion we present examples and/or dig somewhat into these topics to illustrate the implications of assessment illiteracy as well as the knowledge and skills teacher education programs should be promoting among their students.

Examination of the Assessment Literacy Topics

Formative Assessment

The column headed "Formative" is in no way intended to capture the entire instructional process of formative assessment. In 2007, the Council of Chief State School Officers convened a steering committee of national experts and in 2007 published the following definition of formative assessment:

"Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes."

CCSSO and others have subsequently expanded upon this definition, typically citing the key elements in the process:

1. teachers ensuring students understand the learning targets and the criteria for success,
2. teachers gathering rich evidence of student learning by a variety of means (e.g., observation, questioning, quizzes),
3. teachers providing descriptive feedback on gaps in student learning (for teachers to provide the most effective feedback and make the most appropriate adjustments in instruction, they need to be knowledgeable about the learning progressions associated with the learning targets),
4. teachers and students using the feedback to adjust instruction and learning activities,
5. students engaging in self-assessment and meta-cognitive reflection, and
6. teachers activating other students as resources.

The focus here, and in the whole table for that matter, is on evidence gathering and use—steps 2, 3, and 4. Essentially, formative assessment is good teaching. Its components have been researched and proven effective more than any other form of assessment. Therefore, we sincerely believe that teachers should master the entire process, its strategies and techniques, and the important distinction between formative assessment and simply frequent testing. The types of measures (evidence-gathering techniques, tools, activities, and instruments) that can be used in the formative assessment process far exceed those available for any other form of assessment.

To be effective, teachers must have an accurate understanding of the constructs inherent in each standard, unpacking them into learning targets with appropriate attention paid to depth of knowledge (which we'll address in more detail later in the discussion). The quality of the evidence-gathering measures and the quality and use of feedback are matters of teachers' judgment. The concepts of reliability and validity are very important in formative assessment, but not the formal psychometric manifestations of those concepts. A teacher needs to match his or her evidence gathering techniques with a learning target in such a way that they generate rich, conclusive evidence of learning and learning gaps associated with the specific target. There is no measure of the quality of these instructionally embedded techniques—the teacher needs to recognize the quality of evidence using his or her knowledge of the target, understanding of depth of knowledge, and experience.

Perhaps the greatest challenge to effectively implementing formative assessment—as well as most other assessments—is determining what to do next based on the results. What feedback is most appropriate? What instructional adjustments make sense? What type of intervention (from remediation to enrichment) will more effectively promote student growth? (See Goertz, Nabors Olah, & Riggin, 2010 for more information.) Certainly, a sound understanding of the assessment "instrument" and its appropriateness for the intended purpose is essential for using the results to have a positive impact on student learning. However, far more is required, extending into content knowledge, pedagogy, and an understanding of learning progressions. Such elements are beyond the scope of this paper. The objectives/learning targets associated with effectively using assessment results should inform these other components of teacher education programs.

Classroom Summative Assessment

Classroom summative assessment covers a range of applications—from end-of-lesson and -unit tests to end-of-semester or -course tests. The key attribute is that these tests measure what students know and can do after instruction has been completed or at designated milestones in ongoing learning activity. For purposes of classroom summative assessment (the source of grades), a teacher needs to understand the strengths and weaknesses of different types of measures and the consequences of overemphasizing any one. Several well-known options exist, each with its own characteristics and implications for testing time, alignment, reliability, and how results can be used.

For example, while selected-response items take less time, are easy to score (even via automated approaches), and can efficiently gather evidence of whether students grasp basic knowledge and skills, it is more difficult to write them in a way that will effectively measure greater depth of knowledge and higher-order thinking skills. It takes many more selected-response items than open-response items to generate comparably reliable results. In addition, by only revealing students' selected answers and not their work, teachers must infer why students picked the wrong—or even the right—answer. Not all students

selecting a particular answer do so for the reason implied by the answer option itself. Carrying this point one step further, converting a four-option, selected-response item to a short-answer item will generate far, far more than four answers. What happens to all these other answers (and the insights they provide) when students only have four choices?

More performance-based measures, such as open-ended or constructed-response items or performance tasks, take more time and thought to develop and score and require considerable thought in establishing rubrics and scaffolding (this is a good example of how the contents in the Table 1 cells must be unpacked to fully define the assessment literacy domain). On the other hand, they tend to be more rigorous, can more readily tap higher-order thinking skills and, by requiring students to show their work, leave no doubt in teachers' minds about students' learning strengths and weaknesses. This last factor makes them more valid for informing instruction than selected-response items.

In our work, we have identified some harmful myths that many teachers believe about what constitutes good and bad items. For example:

- any selected-response item can and should only measure a single, isolated concept or skill,
- one option in a selected-response item can be a witty throwaway to help keep students entertained and therefore engaged, or
- in creating the wrong options for a multiple-choice item, one should be close to correct and another should be obviously wrong.

Also, some fundamental principles exist that we have often seen ignored in school- or classroom-based tests— for example, the importance of

- avoiding construct-irrelevant factors, such as excessively long and complex language in a mathematics problem, or
- avoiding options in selected-response items that stand out for reasons other than their correctness or incorrectness.

Test quality is a function of many factors, including match to purpose, alignment to standards (content and depth of knowledge), item/task quality, and characteristics of the overall test itself (the test questions as a set). The concepts of reliability and validity are very important here, but still not necessarily at the formal, psychometric level. A teacher needs to use instinct and experience to know that his or her test adequately covers the intended domain and depth of knowledge. Yet, a more formal understanding of the alignment categories listed in the next column of the table would benefit teachers in constructing their own tests (using items they create and/or select from an item bank) with sufficient content validity.

More so with classroom summative assessment than with formative assessment, when interpreting assessment results and contemplating next steps, teachers should ask themselves a simple question: Do the data provide meaningful information about student performance sufficient to act on or do they prompt further questions about student learning? Often, the answer depends upon the application of a fundamental principle: Valid assessment requires multiple sources of evidence collected over time (professional development specialists often refer to this as triangulation of data). Multiple measures provide a solid foundation for analyzing student achievement and for identifying student learning needs—the essential precursor to effective feedback, intervention, instructional adjustments.

In addition, the narrower the focus of a test, the more diagnostic it can be—the more trust you can have in conclusions about a specific strength or weakness of a student or group of students. However, more focused does not mean fewer items. The longer the test, the more reliable it is because it yields more evidence and provides a better sampling of the domain. Understanding the strengths, weaknesses, and appropriates uses of the full range of measures, as well as the principles underlying high quality test creation, is essential to generating reliable and valid results to inform instructional decisions.

Finally, grading practices play an important role in learning. Schafer (1993) identified several grading practices that actually inhibit learning. It has been disappointing in our work to find that many of those practices remain prevalent in our schools today. Historically, for example, teachers have struggled in distinguishing between scoring and grading. Our experiences with educators as part of a Gates Foundation-funded R&D project that has involved evaluating student work, has confirmed this. Teachers tend to inappropriately consider factors beyond actual student work in their scoring. Instead, it is when they translate the scores to grades that they can more

appropriately take such factors as grade level, time of year, etc. into account.

External Interim and Summative Assessment

External interim and summative assessments cover another broad range of instruments intended to address an equally broad range of purposes. They include the full suite of measures (along with their strengths and weaknesses) listed under the Classroom Summative column. External assessments tend to have higher stakes associated with them (often for teachers and administrators—as well as students). Consequently, considerations regarding technical quality take on far greater importance.

Teacher involvement with these assessments varies widely. In some schools and districts, teachers help develop "common" assessments or they might participate in review committees evaluating third-party tests or item banks. Teacher review committees are a standard component when developing statewide assessments, a very formal and rigorous process that addresses alignment in content and depth of knowledge, as well as such considerations as bias and sensitivity. Teachers can—and should—consider bias and sensitivity when creating classroom assessments. Teachers also participate in standard setting activities following the administration of statewide tests. And to varying degrees teachers receive and are expected to appropriately act upon the results of these external assessments. For these and other reasons, teachers need a solid foundation in those elements that relate to the quality and use of these measures.

From a quality perspective teachers need to understand the different types of alignment and the key drivers of reliability and validity. We do not expect teachers to master psychometrics, but they need to understand the fundamental principles, which can—and should—inform all their assessment activities. We can best illustrate the knowledge and skills required for aligning assessments by citing the four alignment types conceptualized by Norman Webb of the University of Wisconsin (Webb, 2007).

- **Categorical Concurrence** means that every item matches a content standard category and subcategory. When creating items, teacher must ensure they match the subject matter and skills they have taught and that are addressed in the standards. A significant challenge arises when teachers use third-party assessment resources. For example, just because an item bank is advertised as being fully "aligned" with a set of standards does not mean that the items in the bank cover every standard well or even every standard at all. Poor coverage or non-coverage of standards by item banks will be more likely as schools transition to the Common Core State Standards.
- **Depth of Knowledge** refers to the level of cognitive complexity of a test item or task. Webb identifies four levels, from Level 1, simple recall of a fact or procedure, to Level 4, extended thinking involving investigation and time to think and process information pertaining to a problem. The language in a standard or a learning target often indicates the highest level intended for instruction and assessment. As noted previously, selected-response items tend to measure lower levels of depth of knowledge, whereas performance-based measures more readily assess greater depth of knowledge.
- **Balance of Representation** refers to the spread of items across content categories and levels of depth of knowledge when a test is intended to cover several categories of content and/or skills. The distribution should be appropriate, either equal or in predetermined proportions, keyed to the standards and curriculum. As an example, one wouldn't want a general achievement measure or interim assessment in mathematics to have all or most of its items from the categories of geometry and measurement. In reading, both literary and informational passages should be included in a general reading comprehension measure. It is usually desirable to have items in a test appropriately cover content categories, as well as represent several levels of cognitive complexity.
- **Range of Knowledge** refers to the coverage of concepts and skills by items in a test within a category, objective, or learning target. For example, if a learning target is "reads and interprets graphs," items should be included that involve direct reading of graphs and require higher-level interpretations of information depicted in graphs. If the target is supposed to include different kinds of graphs (e.g., pictographs, bar graphs, line graphs), then items in a general achievement measure should span those types of graphs, rather than emphasize a single one. Of course, if we're talking about a quiz on a particular type of graph that is being taught at the time, that's a different matter. In other words, there is an appropriate range of knowledge that should be

represented depending on the purpose and target domain of a test.

Among the technical considerations, reliability and validity are perhaps the most important. Reliability, or consistency of measurement, is the extent to which results would be replicated with repeated testing. Validity takes many forms and is strongly related to the intended uses of the assessment results. However, at the most basic level, validity is the extent to which a test is covering the right stuff (knowledge and skills). This is content validity, which is very much related to alignment. These two elements can be heavily steeped in psychometrics, but teachers can—and should—understand the fundamental principles without being overwhelmed with statistics. We previously noted a couple of factors related to these measures:

- by requiring students to show their work, performance-based items are inherently more valid than selected-response items, which are typically indirect measures of the targeted knowledge and skills; and
- there is a direct relationship between the number of items in a test and its reliability, a larger number providing better domain representation (hence, greater validity, too).

Teachers need to understand these principles and their application and can do so without mastering psychometric underpinnings.

It is dismaying to learn how little many educators know about what different kinds of external instruments used for interim and summative purposes can and cannot do. Here are a few examples:

- A district curriculum specialist made significant remediation decisions based on individual student subtest scores from short, efficient commercial tests.
- A school administrator applauded the diagnostic value of a multiple-choice, adaptive test with secure items.
- A teacher using a new online grading tool entered all the quiz and test scores given during the year and somehow, looking at the scores on these vastly different measures, determined whether students demonstrated adequate growth during the year.

Over-interpretation of data from broad, general achievement measures is a problem when local educators try to squeeze diagnostic information for individual students out of measures not designed for that purpose—e.g., fixed-form or adaptive tests designed specifically to generate total test scores as efficiently as possible. The lack of understanding of the meaning of a vendor's claim that an old item bank is now "aligned to the Common Core" is another problem. That each of a set of existing items can be placed in content categories of a new set of standards hardly assures adequate validity or the alignment of a test to the standards.

Underlying the misinterpretation of assessment results relative to group comparisons, growth, etc. is often poor understanding of reporting statistics and comparability. Making comparisons in interpreting assessment results is a necessity, because test scores are only meaningful in relation to something: previous or predicted test scores for a student or particular group of students, other students' test scores, or some established performance standard.

Comparability issues are many, varied, and problematic for educators at all levels—from classroom teachers to university educators to state education policy makers. One of the most common errors made in interpreting test results is making inappropriate comparisons based on measures that are treated as comparable when they are not—for reasons ranging from content coverage to item difficulty. A critical task in every statewide assessment program is a process called equating, which is intended to ensure that reported scores take into account any differences in difficulty between tests from year to year.

Within a classroom, comparability issues may not be significant, as students are often subjected to the same measures. However, across classes, particularly with different teachers teaching the same subject at the same grade level, comparability is often a major concern. How often are common measures across classes or other means of calibration used?

Examples of misinterpretations of test results are exemplified below:

- A teacher team developed pre-and post- tests to monitor student growth without considering comparability, as they intended to subtract raw pre-test scores from raw post-test scores to determine the growth.
- One university reading educator concluded that a state assessment was seriously flawed because it and

an off-the-shelf, commercial reading test produced vastly different percentages of proficient students.

- A state board member couldn't understand why standard setting was done on a state assessment since 70 percent was good enough when he was in school.

Sound reasons exist why each of these examples reflects an incorrect understanding of assessments and results. They illustrate the breadth and depth of issues educators of all stripes create when misinterpreting assessment results. Mastering assessment literacy would avoid them all.

We believe this discussion demonstrates that Table 1 is representative of a broad domain of important content and skills and that the lack of assessment literacy among educators can have dramatic implications, for both teachers and students. Fully defining the domain through a comprehensive set of standards, objectives, and learning targets is a necessary first step toward designing teacher education programs that promote assessment literacy.

We have referred multiple times to the importance of appropriately unpacking the domain of assessment literacy to gain sufficient guidance for the development of pre-service curricula, instructional materials, and practicum experiences, as well as assessments for teacher candidates and practicing teachers. We have noted that some standards and performance measure frameworks partially unpack the domain of assessment literacy. To fully unpack the topics included in Table 1 so that they provide a suitable foundation for promoting and measuring mastery of the domain will require a concerted, inclusive effort involving participants with the requisite expertise. The universal nature of the domain—the ability to apply the content and knowledge across subject areas and grades—would seem to justify undertaking this effort on a collaborative basis.

Table 5 on the following page illustrates how two of the topics included in Table 1—formative assessment and comparability—can be unpacked. We hope it is evident how the specificity of the indicators, can inform program development and performance measure creation.

Pre-Service Promotion of Assessment Literacy

Defining the domain of assessment literacy as described above should form the foundation of an important component of pre-service programs. Such an effort will outline the full scope of knowledge and skills that need to be promoted. We note that every cell in Table 1 calls for both content knowledge and skill in applying that knowledge—in other words, performance. Therefore, educator programs should provide ample opportunities for participants to apply assessment knowledge and skills in real instructional settings, and their performance should be evaluated. We know of a few attempts, including our own preliminary efforts, to create rubrics related to assessment practices. However, they are more limited in scope than what we suggest here.

The question of who is responsible for teaching assessment literacy is an important one. We believe a general opinion exists within teacher education institutions that this responsibility belongs with the reading educators, the mathematics educators, the science educators, etc. There is no question that these individuals should address the unique aspects of assessment within their disciplines. But looking at the broad domain represented in Table 1, it is unlikely that assessment would be well covered in the subject-specific curriculum or methods courses. Furthermore, we have some doubts as to the expertise of many of these educators in the broader domain of assessment literacy. Given the growing importance of assessment and data-driven decision making in our schools, course work focused specifically on assessment literacy topics is needed. To make the content relevant and offer opportunities to apply the knowledge gained through such work, students should be able to undertake subject-specific assignments and activities in their respective areas.

Increasingly, teacher education programs will provide their students with extensive field experience. This could offer invaluable opportunities for teacher candidates to apply their growing knowledge of assessment literacy. With the development of performance rubrics, as suggested above, such authentic work will provide valuable evidence of student progress and inform activities to further build capacity.

Table 5: Examples of Unpacking Assessment Literacy Topics

Category of Measure	Formative
Domain Topic	Use of data to adjust instruction
Standard	Teachers must be able to create/select and effectively use classroom assessments for a variety of purposes.
Objective/ Learning Target	Recognize that formative, qualitative data provides student-by-student and classroom-level information about learning in relation to given learning targets. Effectively use data to adjust instruction, verify instruction, and develop new assessments.
Indicators	Examine student-level data in relation to learning progressions and plans instruction accordingly. Make decisions and gives descriptive feedback based on students' current level of achievement. Synthesize patterns and trends in classroom level data to determine whole class lessons, mini-lessons, small group lessons, individual extensions, and remediation.

Category of Measure	External Interim/Summative
Domain Topic	Comparability
Standard	Teachers and administrators must be able to select and effectively interpret and use test results from external interim and summative assessments designed for a variety of purposes.
Objective/ Learning Target	Recognize that test results are only meaningful in comparison to something (previous test results, predicted test results, results for other groups, or previously established standards of performance) and recognize situations in which comparisons of test scores are appropriate and situations in which they are not.
Indicators	Understand and give examples of why a high correlation between two measures does not mean that they can be used interchangeably. Understand how two tests measuring the same content can be highly correlated, yet produce very different percentages of proficient students. Understand the necessary conditions for pre-test/post-test comparisons. Identify factors, other than the tests themselves, which would render test scores for different groups of students non-comparable (time of testing, administration conditions, testing time, opportunity to learn, etc.)

Pre-Service and In-Service Performance Measures

An adequate definition of the domain of assessment literacy—including objectives and learning targets—should form a solid foundation for determining appropriate performance measures for teacher candidates, as well as practicing teachers. Such measures would help determine assessment literacy mastery for accountability and, especially in the case of teachers, staff development. The instruments could consist of the full range of types of measures, from selected-response and open-response items for evaluating candidates' knowledge of assessment literacy to performance tasks requiring them to construct and administer their own instruments or to interpret sets of assessment results and outline follow-up actions.

Since mastery of assessment literacy is best demonstrated through application—actually using the knowledge and skills in the domain to create and use assessment to promote student learning—we cannot emphasize enough the importance of performance-based measures. As we hope we have illustrated, the domain has many inherent complexities, misinterpretations are rampant, and incorrect practices are far too common. We as a nation are testing more than we ever have, but much of the time, effort, and money invested might be wasted because educators are not getting the information they need to take appropriate actions that promote student learning or because they misuse or misinterpret the information.

Assessments that are composed solely of multiple-choice items, or that merely ask for definitions of terms and concepts, will provide no insights into the ability of candidates or teachers to actually apply the knowledge and skills within the assessment literacy domain. Indeed, for most of the domain, various forms of performance-based assessment, including observation and examination of artifacts, are the best—and we would say the only—effective means of measuring mastery.

In recent years, many groups have promoted much expanded teacher evaluation systems—systems that rely on multiple measures, such as evidence of teacher content and pedagogical knowledge, teacher observations, surveys of students and others, and student achievement gains. Factors such as these are the focus of the ongoing Gates Foundation-funded MET studies. Notably absent from many systems is the evaluation of teaching artifacts: teachers' lesson plans and the tests teachers produce and use, student work they produce, the feedback the teachers provide, and the adjustments they make to their instruction. For example, as noted previously, the Framework for Teaching Evaluation Instrument appears to be solely an observation tool.

The development of criteria for evaluating teachers' assessment practices consistent with the assessment literacy learning objectives and targets, once they have been developed, would be an obvious next step. While we understand the need to include student outcomes in evaluation systems, from the preceding discussion we hope it is now obvious that effectively creating/selecting and using assessment to improve teaching and learning is an ongoing process intimately linked with curriculum, instruction, and the students themselves. Assessment literate teachers effectively implement this process every day. The results at the end of the year take care of themselves. Focusing solely on year-end student test results while ignoring the knowledge and skills inherent in this process—will do little to promote student growth.

The widespread—indeed mandated—focus on educator effectiveness from an accountability perspective has extended to teacher education programs. There is a movement afoot to use the standardized test results of students of teacher education program graduates to evaluate the programs themselves. These efforts typically envision applying a so-called "value-added model" in analyzing the data. The model comes in various forms, but the purpose is to isolate the impact a teacher has on student growth during the year by projecting anticipated student growth on the basis of students' past learning trajectories and comparing it with actual standardized test scores.

The appropriateness of applying this approach to individual teachers as the basis for high-stakes decisions has been hotly debated among measurement experts who have conducted numerous studies over the past several years. In general, the measurement community, cognizant of a range of technical issues, argues against using the results from a single test to make high-stakes decisions. (Clearly, the legislative and regulatory landscape over the past ten years demonstrates that measurement experts don't make public policy.) We share those concerns.

Although we think that student outcomes should be considered in evaluating educators, we believe they should play a different role than many policy makers today endorse. Teacher evaluations should be human judgments by immediate supervisors, informed by a great deal of information, including student test results. We oppose state mandates that student test score gains should be weighted X percent (as high as 50 percent in some states) in the evaluation of teachers. This practice ignores the problems experts have identified with value-added testing, and necessitates quantifying other measures that are not so easily quantified, possibly resulting in student achievement gains counting even more than their assigned weights.

Our reservations about using test scores to gauge the effectiveness of educators diminish when considering them in aggregate in teacher education program evaluation. Most of the measurement issues fall away when applying test scores, such as through a value-added model, to groups of educators for research or formative program evaluation purposes. Therefore, we think these data could be appropriately used in such an application.

However, this is a case in which "inputs" must also be evaluated. From the perspective of this paper, we would want to know what teacher education programs are offering their students to build their capacity in assessment literacy. What courses do they offer and what does the content of those courses include? How much opportunity do students have to apply in class and practica the knowledge gained to create different types of assessments for different purposes and use the results? What performance measures, formative and summative, have been put in place to evaluate and support teacher candidate mastery of assessment literacy and what do these measures cover?

5. Conclusion

Based upon the information presented in the preceding pages, we reiterate our two major conclusions and make corresponding recommendations.

Conclusions:

- In many pre-service programs, the coverage of assessment literacy in course work and practica is incomplete and superficial, leaving graduates unprepared to effectively meet the demands of today's K-12 environment.
- Similarly, the most widely used performance measures' coverage of assessment literacy is incomplete and superficial, rendering it incapable of gauging candidates mastery.

Recommendations:

- Promote candidates' mastery of assessment literacy knowledge and competencies in pre-service programs by including separate course work focused on assessment, embedding assessment topics in content and methods courses, and providing candidates with real-world opportunities to apply what they have learned.
- Flesh out the domain of assessment literacy into a coherent and comprehensive set of objectives and learning targets to provide the specificity needed for designing effective curricula, instructional materials, practica, and formative and summative performance measures.
- Evaluate programs not only in terms of the impact graduates have on student learning, but also on the "inputs," such as the scope and nature of the resources and opportunities devoted to promoting assessment literacy course content, field experiences, and measures—all of which should be heavily performance-based.

We find it remarkable that assessment has attracted so little attention in many pre-service programs and performance measures, despite the fact that it operates as an essential component of the instructional core, accounts for a significant and growing investment of time and resources in K–12 education, and plays an increasing role in high-stakes decisions. Often, assessment is covered incompletely and superficially, leaving new teachers ill prepared to fully promote student learning as they enter the profession. Although high-level standards referring to assessment are available from multiple organizations, minimal if any effort has been devoted to unpacking these standards into specific learning objectives and targets that could inform the design of effective, pre-service programs and performance measures.

Assessment literacy comprises the essential, research-based knowledge, skills, and competencies that all teachers should be prepared to apply in K–12 classrooms—across all grades and content areas. The framework we propose includes a set of assessment topics (knowledge and competencies) in the areas of classroom formative and summative assessment and external interim and summative (including large- scale) assessment. Effective assessment in all its forms is a process carried out by "literate" practitioners. Its performance-based nature means that building capacity and evaluating mastery among teacher candidates, as well as practicing teachers, must similarly be largely performance-based. In general, this has not been the practice.

We envision our proposed framework for the domain of assessment literacy to be a starting point of a process to unpack and flesh out the topics presented in Table 1 into a coherent and comprehensive set of learning objectives and targets. These, in turn, will have sufficient specificity to inform the design and development of high-quality curricula, instructional materials, and practicum experiences to promote assessment literacy among teacher candidates. They can also inform the review of current performance measures for teacher candidates and practicing teachers and serve as the foundation for filling in gaps or creating new or complementary measures. Such measures—used formatively and summatively—could promote capacity building, meaningful evaluation, and continuous improvement. Based upon research and our own practical experience, we think such an effort will raise the level of assessment literacy within the profession and will ultimately improve student outcomes.

We recognize that implementing our recommendations will require a seemingly dramatic change to pre-service programs. Given the scope and complexities of the full domain, a minimally effective pre-service program will require at least one assessment course involving instructors with both measurement expertise

and practical experience. Furthermore, professors who cover content (reading educators, math educators, etc.) will need to embed classroom assessment knowledge and skills in their courses, particularly the set of strategies, techniques, and practices (including formative assessment) unique to their discipline.

Through our research and personal contact with program administrators, we anticipate considerable resistance due to an already over-burdened curriculum and concerns about increased rigor reducing the applicant pool. We also expect objections regarding the use of multiple performance-based measures (from on-demand assessments to observation and evaluation of artifacts) to gauge candidates and practicing teachers' mastery and application of assessment literacy because of the time and effort required. Nevertheless, without implementing such changes, it will be impossible to promote competence in—let alone mastery of—the full domain of assessment literacy and to gauge candidates and practitioners' ability to use assessment to help every student succeed. In short, in the area of assessment literacy and practice, the huge chasm between many pre-service programs and the real-world demands of effective teaching will persist—to the ultimate detriment of our students.

One outstanding question is how to incorporate our recommendations into accreditation standards. We realize that even our high-level framework defining assessment literacy is far too extensive to be included in an accreditation standard. However, just as requirements for appropriate clinical practice might be incorporated in one or more standards, we believe a more process-oriented standard that embodies unpacking the framework and using the results to inform program and performance measure design should be considered. Such a standard should hold programs accountable for (1) the resources and opportunities provided to candidates to build a solid foundation in assessment by the time they graduate, including both course work and practical experience, and (2) the scope and nature of the formative and summative measures programs use to promote mastery and evaluate progress.

Another key question regards implementation of our core recommendations: Who can carry out the necessary work,? While individual institutions could do so on their own, because of the breadth and depth of the assessment literacy domain, most pre-service program staff have neither the knowledge nor the skill to take on this work. Consequently and given the potential widespread use of an unpacked framework and guidelines for performance measures, these activities are conducive to a collaborative effort involving an inclusive group of experts and stakeholders. The outputs could include a valid set of learning objectives and targets as well as item and performance task specifications and samples that could serve as a solid foundation for programs nationwide. The nature of the work products would leave ample room for each program to design its own curriculum, instructional materials, practicum/student learning experiences, and performance measures.

A final question at this point is whether such an effort could have an impact on teacher preparation programs, performance measures, and practitioner development. Perhaps we can learn a few lessons from the experience of the 1990 Standards for Teacher Competence in Educational Assessment of Students. In the spring 2011 issue of Educational Measurement: Issues and Practices, Susan M. Brookhark's article, "Educational Assessment Knowledge and Skills for Teachers" notes the widespread use of the Standards, identifies areas where developments—from formative assessment practice to standards-based assessment—have left the 1990 Standards behind, and suggests updates. Certainly, the standards had backing from three strong organizations and they informed research and teaching practice. Yet, the impact of these fairly specific standards failed to have a significant and lasting impact on teacher education programs and professional practice. Perhaps by gaining broader adoption of the unpacked domain, rather than just standards, we stand a better chance of seeing the day when all educators are assessment literate.

We appreciate the opportunity to submit this paper. We hope it proves useful as efforts proceed to revise the accreditation standards and, ideally, address any gaps in the coverage of assessment literacy.

Bibliography

AdvancED®. (2011). *AdvancED® Standards for Quality Schools*. Retrieved June 18, 2012, from Accredidation: http://www.advanc-ed.org/new-standards-quality

American Federation of Teachers, National Council on Measurement in Education, National Education Association. (1990). *Standards for Teacher Competence in Educational Assessment of Students*. Retrieved November 30, 2012, from http://buros.org/standards-teacher-competence-educational-assessment-students

Black, P., & Wiliam, D. (1998a). *Assessment and Classroom Learning. Assessment in Education: Principles, Policy & Practice*, 5 (1), 7–74.

Black, P., & Wiliam, D. (1998b). Inside the Black Box. *Phi Delta Kappen*, 139–148.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for Learning: putting it into practice*. New York: Open University Press.

Brookhart, S.M. (2011). Educational Assessment Knowledge and Skills for Teachers. *Educational Measurement: Issues and Practices*, 30(1), 3-12.

CCSSO's InTASC. (2011, April). *InTASC Model Core Teaching Standards: A Resource for State Dialogue*. Retrieved June 18, 2012, from Resources: http://www.ccsso.org/Resources/Publications/InTASC_Model_Core_Teaching_Standards_A_Resource_for_State_Dialogue_(April_2011)-x1025.html

Danielson, C. (2011). *2011 Framework for Teaching Evaluation Instrument*. Retrieved July 26, 2012, from The Framework: http://www.danielsongroup.org/article.aspx?page=frameworkforteaching

Educational Testing Services. (2011). *Praxis II: Principles of Learning and Teacher, K–6, 5–8, 7–12*. Retrieved July 7, 2012, from Preparation Materials: http://www.ets.org/praxis/prepare/materials

Goertz, M. E., Nabors Olah, L., & Riggin, M. (2009). *From Testing to Teaching: The Use of Interim Assessments in Classroom Instruction (RR–65)*. Retrieved July 26, 2012, from Research Reports: http://www.cpre.org/research-reports

International Association for K-12 Online Learning. (October 20, 2011). *National Standards for Quality Online Teaching, Version 2*. Retrieved September 19, 2012 from http://www.inacol.org/research/nationalstandards/iNACOL_TeachingStandardsv2.pdf

IRA. (2010). *Professional Standards 2010*. Retrieved June 25, 2012, from Standards 2010: 3 Assessment and Evlautaion: http://www.reading.org/General/CurrentResearch/Standards/ProfessionalStandards2010/ProfessionalStandards2010_Standard3.aspx

Kahl, S. R. (2010, Fall). *Something Old, Something New: What Teachers as Assessors Must Know and Be Able to Do*. Retrieved June 26, 2012, from AdvancEd Source: http://www.advanc-ed.org/advanced-source

Massachusetts Department of Elementary and Secondary Education. (2009, January 1). *Massachusetts Test for Educator Licensure (R)*. Retrieved July 25, 2012, from MTEL Test Objectives: http://www.mtel.nesinc.com/MA_testobjectives.asp

Measured Progress & ETS Collaborative. (2012, April 12). *Smarter Balanced Assessment Consortium General Item Specifications*. Retrieved July 26, 2012, from Smarter Balanced Assessments: http://www.smarterbalanced.org/smarter-balanced-assessments/

Michigan Assessment Consortium. (September 2012). *Assessment Literacy Standards*, Version 3.1.

National Research Council. (2001). *Knowing What Students Know: The science and design of educational assessment*. (J. W. Pellegrino, R. Glaser, & N. Chudowsky, Eds.) Washington, DC: National Academy Press.

NBPTS. (2012). *Literacy: Reading-Language Arts/Early and Middle Childhood*. Retrieved June 23, 2012, from Standards by Certificate: http://www.nbpts.org/the_standards/standards_by_cert?ID=23&x=56&y=12

NCATE. (2008, February). *Professional Standards for the Accredidation of Teacher Preparation Institutions*. Retrieved June 20, 2012, from NCATE Unit Standards: http://www.ncate.org/Standards/NCATEUnitStandards/tabid/123/Default.aspx

NCTM. (2011, April). *NCTM NCATE Draft Standards: DRAFT Middle Level Mathematics*. Retrieved June 20, 2012, from Standards and Focal Points: http://www.nctm.org/NCATEdraft/

Nitko, A., & Brookhart, S. (2011). *Educational Assessment of Students* (6th Edition). Boston, Massachusetts: Pearson Education, Inc., publishing as Allyn & Bacon.

Pearson. (2012). *About the TPA: Teacher Performance Assessment*. Retrieved July 13, 2012, from Teacher Performance Assessment: http://tpafieldtest.nesinc.com/PageView.aspx?f=GEN_AbouttheTests.html

Pearson Education, Inc. (2011a). *Assessment for Professional Knowledge-Elementary Test Framework*. Retrieved July 22, 2012, from National Evaluation Series (TM): http://www.nestest.com/TestView.aspx?f=HTML_FRAG/NT051_PrepMaterials.html

Pearson Education, Inc. (2011b). *MTTC Test Objectives*. Retrieved July 25, 2012, from Michigan Test for Teacher Certificaton: http://www.mttc.nesinc.com/MI_viewFW_opener.asp

Pearson for submission under contract with the NBPTS®. (2012). *National Board for Professional Teaching Standards*. Retrieved June 25, 2012, from Literacy: Reading-Language Arts/Early and Middle Childhood: Scoring Guide for Candidates: http://www.nbpts.org/for_candidates/scoring?ID=23&x=74&y=12

Popham, W. J. (2003). *Test Better, Teach Better: The Instructional Role of Assessment*. Alexandria: ASCD.

Rhode Island Department of Elementary and Secondary Education. (2011, November 11). *Regulations Governing the Certification of Educators in Rhode Island*. Retrieved July 16, 2012, from Office of Educator Quality and Certification: http://www.ride.ri.gov/educatorquality/certification/

Shafer, W. D. (1993). Assessment Literacy for Teachers. *Theory into Practice*, 32 (2), 188–126.

Shepard, L., Hammerness, K., Darling-Hammond, L., Rust, F., Snowden, J. B., Gordon, E., et al. (2005). Assessment. In L. Darling-Hammond, & J. Bransford, *Preparing Teachers for a Changing World* (pp. 275–326). San Fancisco: Jossey-Bass.

Stiggins, R. (2007). Assessment through students eyes. *Education Leadership*, 64 (8), 22–26.

Webb, N. (2007, September). *Aligning Assessments and Standards*. Retrieved July 26, 2012, from Wisconsin Center for Education Research: http://www.wcer.wisc.edu/news/coverstories/aligning_assessments_and_standards.php

Wood, G. H., Darling-Hammond, L., Neill, M., & Roschewski, P. (2007, August 27). *Refocusing Accountability: Using Local Performance Assessments to Enhance Teaching and Learning for Higher Order Skills*. Retrieved July 26, 2012, from Fair Test: http://www.fairtest.org/refocusing-accountability-using-local-performance-